Protonated water clusters by stochastic approaches : probing machine learning resilience against quantum Monte Carlo noise

Matteo Peria



Thèse de doctorat de Sorbonne Université

École Doctorale de Physique en Île-de-France - ED 564

Réalisée à

l'**Institut de minéralogie, de physique des matériaux et de cosmochimie**

Sujet de la thèse:

Protonated water clusters by stochastic approaches: probing machine learning resilience against quantum Monte Carlo noise

présentée par

Matteo PERIA

Directeurs de thèse: Michele Casula et A. Marco Saitta

Soutenue publiquement à Paris le 12 novembre 2024 devant le jury composé de :

Bertrand LAFORGE	Professeur	Sorbonne Université	Président
Thierry DEUTSCH	Directeur de recherche	CEA Grenoble	Rapporteur
Matthias RUPP	Directeur de recherche	LIST Luxembourg	Rapporteur
Federica AGOSTINI	Maître de conférences	Université Paris Saclay	Examinatrice
Rocio SEMINO	Maître de conférences	Sorbonne Université	Examinatrice
Michele CASULA	Directeur de recherche	CNRS, Sorbonne Université	Directeur de thèse
A. Marco SAITTA	Professeur	Sorbonne Université	Directeur de thèse

Abstract

A complete understanding of the hydrogen bond and proton transfer mechanism in water is still lacking, since it requires an accurate potential energy surface (PES) and very expensive quantum mechanical simulations of the nuclear part. Protonated water clusters are useful building blocks to study the proton hopping dynamics, which we simulate here in the protonated water hexamer $H^+(H_2O)_6$ by a combination of state-of-the-art quantum Monte Carlo (QMC) methods and path-integral Langevin dynamics (PILD). We report a remarkably low thermal expansion of the hydrogen bond from zero up to 300 K, after which the hydrogen bond strength weakens. This behaviour is explained by proton delocalisation, which is favoured by the synergy of nuclear quantum effects and thermal activation, making the near-room-temperature range of 250K-300K optimal for proton transfer. In the second part of this work we test if machine learning interatomic potentials (MLIPs), based on kernel methods or on neural networks, can reproduce the PES of protonated water clusters that would be infeasible to simulate with current high-level computational chemistry methods, either in size or in duration of the simulation. The QMC+PILD learning approach yields very accurate results, which are however affected by the intrinsic noise inherent in the stochastic sampling of both nuclear and electronic phase space. We prove that the QMC noise is not necessarily detrimental to the learning of energies and forces and we determine under which conditions one can derive accurate and reliable MLIPs from QMC data.

Résumé

Une compréhension complète des mécanismes qui gouvernent la liaison hydrogène et le transfert de proton dans l'eau fait encore défaut. Une difficulté majeure qui entrave notre compréhension de ces phénomènes est représentée par le temps de calcul important nécessaire à modéliser les processus en jeu, nécessitant une surface d'énergie potentielle (PES) précise et un traitement quantique à la fois des électrons et des noyaux. Dans ce cadre, les clusters d'eau protonée sont des briques utiles pour étudier la dynamique des sauts de proton, car leur taille finie les rend plus simples à traiter que l'eau liquide. Dans cette thèse, nous avons analysé les résultats sur l'hexamère d'eau protonée $H^+(H_2O)_6$ obtenus en combinant les méthodes de Monte Carlo quantique (QMC) les plus avancées et la dynamique de Langevin par intégrales de chemin (PILD). Nous avons découvert une expansion thermique de la liaison hydrogène remarquablement faible de zéro jusqu'à 300 K, température après laquelle la liaison hydrogène devient moins forte. Ce comportement s'explique par la délocalisation du proton, favorisée par la synergie entre effets quantiques nucléaires et activation thermique, ce qui rend la plage des températures optimales pour le transfert de proton proche de celle ambiante (250K-300K). Dans la deuxième partie de ce travail, nous avons vérifié que les potentiels interatomiques d'apprentissage automatique (MLIP), basés sur des méthodes à noyau (kernel methods) ou sur des réseaux de neurones, peuvent reproduire le PES des clusters d'eau protonés. Leur dynamique serait impossible à reproduire avec les méthodes les plus précises de chimie théorique, à la fois en termes de taille et de durée de la simulation. En revanche, l'approche d'apprentissage basée sur les données QMC+PILD donne des résultats très précis, qui sont toutefois affectés par le bruit intrinsèque de l'échantillonnage stochastique de l'espace des phases nucléaire et électronique. Nous montrons cependant que le bruit QMC n'est pas préjudiciable à l'apprentissage automatique des énergies et des forces et nous déterminons les conditions auxquelles on peut générer des potentiels MLIP fiables en partant des données QMC.

Contents

Acronyms

ACE Atomic Cluster Expansion. 80–82, 84, 85, 88, 104
ACSFs Atom Centered Symmetry Functions. 80, 81, 87, 88, 101
AIMD *ab initio* molecular dynamics. 3, 8, 10, 56, 71

BO Born-Oppenheimer. 8, 10, 32, 43, 46
BOMD Born-Oppenheimer molecular dynamics. 10

CC Coupled Cluster. 15, 29, 73, 74, 104, 105, 112
CI Configuration Interaction. 15
CPMD Car-Parrinello molecular dynamics. 10, 112

DFT Density Functional Theory. 16, 18, 19, 22, 51, 104, 105, 112, 134

FCHL19 Faber Christensen Huang Lilienfeld. 80, 81, 84, 88, 89
FDT Fluctuation-Dissipation Theorem. 35
FPE Fokker-Planck equation. 36

GNN Graph neural network. 102, 103
GPR Gaussian process regression. 94, 135

H-bond Hydrogen bond. 1–3, 10, 18, 41–43, 53, 54, 56, 57, 66, 142
HDNN High-dimensional neural network. 100, 101, 104, 105

KRR Kernel ridge regression. 90, 121, 135

MACE Message-passing Atomic Cluster Expansion. 78, 104, 139, 140
MBE Many-body expansion. 73–75, 104, 108, 113, 118, 130, 131, 133, 139
MD Molecular Dynamics. 3, 32, 35, 53, 56, 59, 61, 63, 67, 108, 113–115, 117, 127, 128, 132, 134
ML Machine Learning. 4, 77, 78, 93, 100, 104, 105
MLIP Machine learning interatomic potential. 1, 2, 4, 79, 80, 100, 104, 105, 107–109, 111, 113, 118–120, 123, 127, 130–133
MP Møller-Plesset perturbation theory. 15, 55, 73, 74, 104, 105, 112

MPNN Message-passing neural network. 102–104, 120

NN Neural network. 96, 98–101, 103
NQEs Nuclear Quantum Effects. 3, 41–43, 53, 54, 56–58, 62–66, 108, 112, 113, 139, 141

OQML Operator Quantum Machine Learning. 78, 95, 120, 121, 123, 126, 127, 130–132, 135, 139, 140

PCF Pair Correlation Function. 56, 58, 128, 135, 137
PES Potential Energy Surface. 1, 4, 8, 10, 32, 41–43, 46, 51, 54–56, 61, 71, 73, 75, 77, 78, 80, 92, 94, 99, 107, 108, 111–113, 115, 127, 128, 132, 134, 135, 139–142
PILD Path integral Langevin dynamics. 47, 67, 132
PILE Path integral Langevin equation. 48, 50
PIMC Path integral Monte Carlo. 42
PIMD Path integral molecular dynamics. 41–43, 59, 63–67
PIOUD Path integral Ornstein-Uhlenbeck process. 41, 48–50
PIPs Permutationally-invariant polynomials. 74, 79, 80, 99, 113
PT Proton Transfer. 1, 2, 4, 41, 55, 56, 63–66, 139

QMC Quantum Monte Carlo. 1–4, 19, 22, 23, 27, 28, 31, 37, 39, 51, 53, 56, 60, 61, 63, 64, 66, 67, 105, 107–110, 112–119, 123, 126, 132, 134, 139–142

RPMD Ring polymer molecular dynamics. 45, 51, 64, 108, 113, 130, 135

SDE Stochastic Differential Equation. 35, 36
SOAP Smooth Overlap of Atomic Positions. 80, 114

VMC Variational Monte Carlo. 19, 22, 23, 28, 29, 55, 56, 59, 61, 67, 109, 110, 113

ZPE Zero Point Energy. 3, 41–43

List of Figures

List of Tables

Introduction

The behaviour of the proton in water has long puzzled chemists and physicists, leading to centuries of debate about all of its aspects [1–7]. This complexity is distilled into a simple chemical expression:

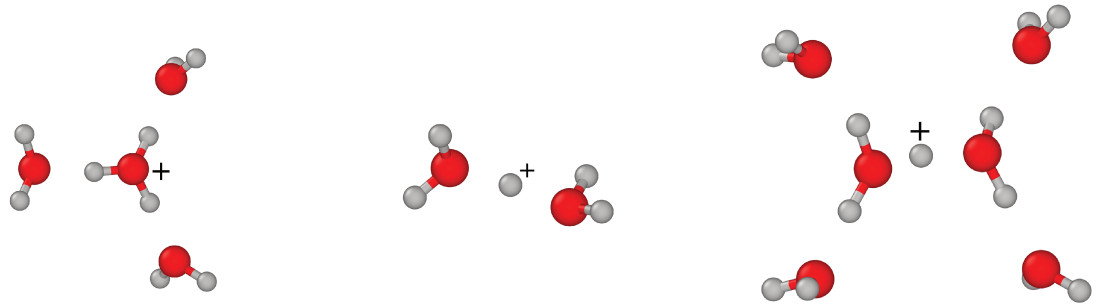
$$H^+(aq). \tag{1}$$

This seemingly straightforward formula [8, 9] conceals a rich history of scientific inquiry into how the proton is hydrated and diffuses in bulk water through the proton transfer (PT) mechanism [10]. To this end, the structure and dynamics of the hydrated proton have been explored by a wealth of experiments, supported by countless theoretical models and numerical simulations [11, 12].

Molecular simulations [13], which provide full control over the accuracy of interatomic interactions, offer a unique and detailed view of the proton jumping from one molecule to another. However, since PT occurs over multiple length and time scales—from the hopping frequency across adjacent water molecules to the breaking of a Hydrogen bond (H-bond) in second solvation shell, followed by a rearrangement of the whole structure around the proton [14, 15]—there is often a trade-off between the accuracy of computational methods and the size and duration of the simulation.

The goal of this thesis is twofold. First, we apply advanced methods that fully account for the quantum nature of both the electrons and nuclei. Specifically, we use Quantum Monte Carlo (QMC) as electronic structure method to derive the potential energy surface (PES) necessary to drive the dynamics of the nuclei, described within the path integral formalism. Given the computational cost of these highly accurate techniques, we focus on the protonated water hexamer, $H^+(H_2O)_6$, to study the temperature effects on proton hydration and hopping.

The second objective is to bridge the gap between accuracy and the limitations imposed by system size and simulation time. Over the last two decades, machine learning interatomic potentials (MLIPs) have emerged as a tool that can reproduce the results of advanced electronic structure calculations at a fraction of the cost. Here, we test on the Zundel cation $H_5O_2^+$ whether MLIPs can learn energies and forces derived from stochastic methods like QMC, providing a stable and reliable potential energy surfaces on top of which we can run extended simulations.



(a) Eigen ion (b) Zundel ion (c) Protonated water hexamer

Figure I.1: **Protonated water clusters considered in this work.** (a) the Eigen ion $H_9O_4^+$ is a fully solvated hydronium H_3O^+, (b) the Zundel ion $H_5O_2^+$ is a solvated proton equally shared between two water molecules, and (c) the protonated water hexamer $H^+(H_2O)_6$ is the smallest cluster which includes both forms of proton solvation.

The hydrated proton computational dilemma: system size vs. accuracy

There are two approaches when it comes to simulating the hydrated proton: either it is embedded in the bulk, modeled using periodic boundary conditions, or it is solvated in a finite-sized water cluster, $H^+(H_2O)_n$. The former approach is ideal for gaining a global understanding of the Grotthuss mechanism [1], which describes the sequence of PT reactions between water molecules. This process can be viewed as a series of proton hops across the H-bond network. Given the length scale of this phenomenon, which can affect multiple molecules along a "water wire" [16], computational efficiency is prioritized over precision in the electronic description.

On the other hand, protonated water clusters serve as the building blocks for our understanding of the hydrated proton, able to replicate key motifs and structures also found in the bulk while maintaining high accuracy in the description of both electrons and nuclei. In fact, the first attempts to explain PT were made by considering the two preferred solvation complexes of the proton in water: the Eigen ion, $H_9O_4^+$ [17], and the Zundel ion, $H_5O_2^+$ [18], as shown in Fig. I.1a and Fig. I.1b, respectively. It did not take long for the computational community, already engaged in water simulations since the early days of computer modeling [19, 20], to address the challenges of determining the optimal geometry of solvated hydronium [21], as well as the proton dynamics in the Zundel complex [22, 23].

Thesis outline

These pioneering studies on protonated water clusters were followed by many others, which will be reviewed in subsequent Chapters. In the first part of the thesis we continue this line of research by investigating the protonated water hexamer, $H^+(H_2O)_6$, or $H_{13}O_6^+$, shown in Fig. I.1c. It has been confirmed that both Eigen and Zundel complexes coexist in the hexamer [24–26], which appears to be the smallest one able to exhibit this feature [12]. It follows that

this system is suitable to reproduce the isomerization process underlying the proton transfer.

In Chapter 1, we briefly introduce *ab initio* molecular dynamics (AIMD), a computer simulation method where the system configurations are sampled by iteratively solving the equations of motion using forces derived from quantum theory. Unlike empirical potentials with predefined functional forms, AIMD significantly enhances the predictive power of our simulations, allowing us to conduct detailed "in silico" experiments that would otherwise be unattainable.

Proton transfer occurs within the water matrix, an incredibly complex environment characterized by multiple types of interactions. The strength of H-bonds is comparable to that of covalent bonds, blurring the line between the two. The presence of a charged species also requires careful consideration of long-range interactions, including Coulomb and van der Waals forces. Additionally, the typical timescale for proton transport in water under ambient conditions is around 1 picosecond [27], indicating a relatively small activation barrier. These considerations lead us to the topic of Chapter 2, where we address the need for highly accurate electronic structure methods, such as quantum Monte Carlo.

Once the forces from electronic calculations are determined, or estimated, it is the nuclei's turn to move through molecular dynamics (MD), the focus of Chapter 3. Many experimental findings on water suggest that nuclei, particularly lighter ones, should be treated quantum mechanically. Nuclear quantum effects (NQEs) are especially relevant for hydrogens and protons, which can exhibit zero-point energy (ZPE), proton delocalization, energy discretization, and proton tunneling [28, 29]. NQEs are significant in water [30–32], and in this Chapter we describe the MD schemes that allow to treat them, together with the noisy forces coming from QMC calculations.

In Chapter 4, we present our results from applying the above methods to study the protonated water hexamer, the largest system that can be studied by path integral molecular dynamics and quantum Monte Carlo simulations with current computational means.

To extend our results to larger time and length scales, we must turn to potentials, the subject of the second part of this thesis.

After discussing the advantages and limitations of various water potentials in Chapter 5, it becomes clear that the primary challenge lies in capturing the complex variety of interactions in water and modeling their quantum nature with predefined functional forms. The interpolation of water's PES and its extrapolation to larger clusters or to longer simulations is an attractive approach for better understanding PT. However, high dimensionality, nonlinearity, and the noise affecting QMC-PES present significant challenges.

These considerations lead us to Chapter 6, which is devoted to machine learning interatomic potentials (MLIPs). In recent years, machine learning (ML) methods have become increasingly popular for solving high-dimensional regression problems across various disciplines, and chemistry and physics are no exception. Machine Learning refers to a broad range of techniques designed to find meaningful patterns from a given dataset that forms the "experience" of the learner, whether in classification tasks (assigning labels to data) or regression tasks (predicting

continuous values). The strength of these algorithms lies in their adaptability, hence the expression *data-driven modeling*. Instead of assuming an expected functional form underlying the data, these methods allow the learner to adapt based on the input data, making them highly flexible and powerful.

Although MLIPs are widespread in the modeling community, little is known about their ability to interpolate noisy data and reproduce reliable PES—an essential requirement for extending the findings of QMC-MD simulations. In Chapter 7 we conduct a comprehensive study on the robustness of MLIPs in learning noisy PES estimated with stochastic electronic structure methods. We apply well-established ML testing methods and rigorously compare the MLIP-derived physics with the one obtained from *ab initio* calculations.

Finally, in Chapter 8, we summarize our main conclusions and outline potential directions for future research.

Part I

Quantum Monte Carlo driven ring polymer molecular dynamics

Chapter 1

Ab initio molecular dynamics

Atoms, molecules and condensed matter systems can be described as a collection of interacting nuclei and electrons that require a quantum mechanical framework for understanding their physical properties.

Consider a system of M nuclei with masses $\{m_a\}_{a=1,\cdots,M}$ and charges $\{eZ_a\}_{a=1,\cdots,M}$, and N electrons with mass m and charge e; let $\{(\mathbf{q}_a, \mathbf{p}_a\}_{a=1,\cdots,M}$ denote the positions and momenta of the nuclei, and $\{(\mathbf{r}_i, \boldsymbol{\pi}_i\}_{i=1,\cdots,N}$ those of the electrons; to simplify the notation we will represent the nuclear and electronic positions as single vectors, $\mathbf{q} = \{\mathbf{q}_a\}_{a=1,\cdots,M}$ and $\mathbf{r} = \{\mathbf{r}_i\}_{i=1,\cdots,N}$, respectively. The time evolution of such a compound follows the spin-free non-relativistic time-dependent Schrödinger equation [33]:

$$i\hbar\frac{\partial}{\partial t}\Psi\left(\mathbf{r},\mathbf{q},t\right) = \hat{H}\Psi\left(\mathbf{r},\mathbf{q},t\right), \tag{1.1}$$

where $\Psi(\mathbf{r},\mathbf{q},t)$ is the total wavefunction and $\hat{H}$ is the Hamiltonian operator, which we can obtain from the classical Hamiltonian $H(\mathbf{r},\mathbf{q})$ of M positively charged particles interacting with N negatively charged ones:

$$H(\mathbf{r},\mathbf{q}) = \sum_{i=1}^{N}\frac{\boldsymbol{\pi}_i^2}{2m} + \sum_{a=1}^{M}\frac{\mathbf{p}_a^2}{2m_a} + \frac{1}{2}\sum_{i,j}^{N}\frac{e^2}{|\mathbf{r}_i-\mathbf{r}_j|} + \frac{1}{2}\sum_{a,b}^{M}\frac{Z_aZ_be^2}{|\mathbf{q}_a-\mathbf{q}_b|} - \frac{1}{2}\sum_{i,a}^{MN}\frac{Z_ae^2}{|\mathbf{r}_i-\mathbf{q}_a|}. \tag{1.2}$$

By replacing the momenta with their respective quantum operators in the position representation, $\hat{\mathbf{p}}_a = -i\hbar\nabla_{\mathbf{r}_a} = -i\hbar\nabla_a$ and $\hat{\boldsymbol{\pi}}_i = -i\hbar\nabla_{\mathbf{r}_i} = -i\hbar\nabla_i$, we obtain:

$$\hat{H} = \underbrace{-\sum_{a=1}^{M}\frac{\hbar^2}{2m_a}\nabla_a^2}_{\hat{T}_\mathrm{n}} \underbrace{-\frac{\hbar^2}{2m}\sum_{i=1}^{N}\nabla_i^2}_{\hat{T}_\mathrm{e}} + \underbrace{\frac{1}{2}\sum_{i,j}^{N}\frac{e^2}{|\mathbf{r}_i-\mathbf{r}_j|}}_{\hat{V}_\mathrm{ee}} + \underbrace{\frac{1}{2}\sum_{a,b}^{M}\frac{Z_aZ_be^2}{|\mathbf{q}_a-\mathbf{q}_b|}}_{\hat{V}_\mathrm{nn}} \underbrace{-\frac{1}{2}\sum_{i,a}^{MN}\frac{Z_ae^2}{|\mathbf{r}_i-\mathbf{q}_a|}}_{\hat{V}_\mathrm{en}}, \tag{1.3}$$

where we dubbed the nuclear kinetic energy $\hat{T}_\mathrm{n}$, the electronic kinetic energy $\hat{T}_e$, and we defined the remaining Coulombic interaction terms $\hat{V}_\mathrm{ee}$, $\hat{V}_\mathrm{nn}$ and $\hat{V}_\mathrm{en}$ for later convenience.

The Hamiltonian does not depend on time, consequently Eq. 1.1 can be simplified into an eigenvalue problem by assuming stationary states:

$$\hat{H}\Psi(\mathbf{r},\mathbf{q}) = E\Psi(\mathbf{r},\mathbf{q}). \tag{1.4}$$

Still, the problem posed by Eq. 1.4 is not exactly solvable, therefore one must resort to approximate methods. These are collectively known as *ab initio* molecular dynamics (AIMD), meaning molecular dynamics 'from first principles', because they are based on pure quantum theory alone, without relying on experimental data or *ad hoc*, though physically motivated, parametrizations.

Different AIMD methods are defined by increasing levels of approximation. Since the result of such procedure, namely the potential energy surface (PES), is a key object of this thesis, we will spend a few lines on its derivation in Section 1.1, and we will explain how the concept of PES has been exploited in the realm of quantum simulations of water in Section 1.2.

1.1 Born-Oppenheimer approximation

The first historical approximation from Born and Oppenheimer [34] is actually based on the classical assumption that the nuclei are not far from equilibrium and the nuclear kinetic energy T_n is small enough to be treated as a perturbation of the electronic Hamiltonian H_e, owing to the large mass difference between electrons and nuclei[1]

$$\hat{H}_{\mathrm{e}} = \hat{T}_{\mathrm{e}} + \hat{V}_{\mathrm{ee}} + \hat{V}_{\mathrm{nn}} + \hat{V}_{\mathrm{en}}. \tag{1.5}$$

This is enough to justify an expansion of the quantum eigenvalue problem $\hat{H}\Psi(\mathbf{r};\mathbf{q}) = E\Psi(\mathbf{r};\mathbf{q})$ with respect to a power of the electron-nuclei mass ratio m_e/M_0, where M_0 is either one of the nuclear masses or their mean. In the original paper the choice of the expansion parameter was $\kappa^4 = m_e/M_0$, but other choices are possible, for example in the mathematical physics community $\kappa^2 = m_e/M_0$ is more common. In either case, the important point proved in the original paper is that only even powers of κ contribute to the energy. If we consider the latter convention up to the second order, we arrive to

$$\hat{H} = \hat{H}_{\mathrm{e}} + \kappa^2 \hat{H}_{\mathrm{n}}, \tag{1.6}$$

with $\hat{H}_{\mathrm{n}} = \kappa^{1/2}\hat{T}_{\mathrm{n}}$. The first term is the electronic energy, the second can be related to ionic vibrations. Eventually, a quartic order term would describe the ionic rotational energy, and higher order terms the coupling between the previous ones. Then, for several configuration of the nuclei, $\mathbf{q}$, and in the limit of $\kappa \to 0$, it is assumed that it is possible to find the solution of the eigenvalue problem of the unperturbed electronic Hamiltonian H_e. Since sending κ to zero means that the nuclear kinetic energy vanishes, therefore the nuclei are fixed, H_e is also called clamped-nuclei Hamiltonian. The set of eigenfunctions of H_{e} is used to calculate the full electrons-ions wavefunction $\Psi(\mathbf{r},\mathbf{q},t)$ and its corresponding eigenvalue E, by simple product with a nuclear wavefunction [36].

$$\Psi(\mathbf{r},\mathbf{q},t) \approx \Phi(\mathbf{r};\mathbf{q})\Omega(\mathbf{q},t). \tag{1.7}$$

[1] It can be proved by energy argument that close to the ground state energy E_0 the kinetic term T_n is automatically small, without the need of exploiting the mass ratio. However, the latter is still necessary for a general proof of the Born-Oppenheimer (BO) approximation that includes also excited states [35].

We start by introducing an extension of Eq. (1.7), for the exact form of $\Psi(\mathbf{r},\mathbf{q},t)$, solution of Eq. (1.1). This ansatz, named Born-Huang, comprises the sum of several terms, and it is still based on the factorization of an electronic and a nuclear part [37, 38]. Following [39] we introduce the Born-Huang ansatz:

$$\Psi(\mathbf{r},\mathbf{q},t) = \sum_l \Phi_l(\mathbf{r};\mathbf{q})\Omega_l(\mathbf{q},t), \tag{1.8}$$

where $\Phi_l(\mathbf{r};\mathbf{q})$ are orthonormal eigenfunctions of the time-independent electronic Schrödinger equation for the clamped-nuclei Hamiltonian $\hat{H}_\mathrm{e}$

$$\hat{H}_\mathrm{e}\Phi_l(\mathbf{r};\mathbf{q}) = E_l(\mathbf{q})\Phi_l(\mathbf{r};\mathbf{q}), \tag{1.9}$$

and, as such, they span the space of the electronic degrees of freedom for fixed nuclei[2], with $\mathbf{q}$ treated as a parameter. Instead the nuclear wavefunctions $\Omega_l(\mathbf{q},t)$ are described by functions that are neither orthonormal nor normalized [40]. Inserting this ansatz in Eq. 1.1, followed by multiplication on the left by the single adiabatic state $\Phi_k^*(\mathbf{r};\mathbf{q})$ and integration over the electronic degrees of freedom $\mathbf{r}$ bring us to

$$i\hbar\frac{\partial\Omega_k(\mathbf{q},t)}{\partial t} = \left[-\sum_a \frac{\hbar^2}{2m_a}\nabla_a^2 + E_k(\mathbf{q})\right]\Omega_k(\mathbf{q},t) + \sum_l C_{kl}\Omega_b(\mathbf{q},t), \tag{1.10}$$

where the non-adiabatic coupling operator C_{kl} is a short notation for

$$C_{kl} = \int \mathrm{d}\mathbf{r}\Phi_k^*(\mathbf{r};\mathbf{q})\left[-\sum_a \frac{\hbar^2}{2m_a}\nabla_a^2\right]\Phi_l(\mathbf{r};\mathbf{q}) + \sum_a \frac{1}{m_a}\left[\int \mathrm{d}\mathbf{r}\Phi_k^*(\mathbf{r};\mathbf{q})(-i\hbar\nabla_a)\Phi_l(\mathbf{r};\mathbf{q})\right][-i\hbar\nabla_a]. \tag{1.11}$$

Equations 1.10 and 1.11 tell us that the nuclear wavefunctions $\Omega_k(\mathbf{q},t)$ evolve following the adiabatic potential energy surface $E_k(\mathbf{q})$, for which the diagonal elements C_{kk} represents a small correction; eventually the nuclei can hop from one electronic state to another according to the off-diagonal terms of the nonadiabatic coupling C_{kl}. This picture comprises several potential energy surfaces, $\{E_k(\mathbf{q})\}$.

The adiabatic approximation consists in neglecting the off-diagonal contributions of the coupling matrix, simplifying Eq. 1.10 to

$$i\hbar\frac{\partial\Omega_k(\mathbf{q},t)}{\partial t} = \left[-\sum_a \frac{\hbar^2}{2m_a}\nabla_a^2 + E_k(\mathbf{q}) + C_{kk}(\mathbf{q})\right]\Omega_k(\mathbf{q},t), \tag{1.12}$$

which means that the quantum state k of the electrons never changes during the dynamics. Instead it adapts parametrically to the slow nuclear degrees of freedom. This would be equivalent to inserting in the original Eq. 1.1 the single product state

$$\Psi(\mathbf{r},\mathbf{q},t) = \Phi_k(\mathbf{r},\mathbf{q})\Omega_k(\mathbf{q},t). \tag{1.13}$$

[2]The sum is done as if all the eigenfunctions of the basis were discrete, but actually one should either count also contributions from the continuous spectrum, which is tricky, or approximate the total wavefunction on a restricted set of eigenfunctions [35].

A further step bring us to the original Born-Oppenheimer (BO) approximation, where even the correction C_{kk} is neglected:

$$i\hbar\frac{\partial\Omega_k(\mathbf{q},t)}{\partial t} = \left[-\sum_a \frac{\hbar^2}{2m_a}\nabla_a^2 + E_k(\mathbf{q})\right]\Omega_k(\mathbf{q},t). \tag{1.14}$$

For almost a century the BO approximation has proven to be a great tool to interpret chemistry concepts in the light of quantum mechanics. Also, it has been of paramount importantance in the development of computational quantum chemistry, because it allows one to study the dynamics of a system relying solely on a single potential energy surface, most often the electronic ground state one, $E_0(\mathbf{q})$. This approximation is at the basis of Born-Oppenheimer molecular dynamics (BOMD), which is one of the most widespread flavour of AIMD, together with Car-Parrinello molecular dynamics [41].

1.2 AIMD of protonated water

Simple water models cannot capture its multifaceted behaviour, from the challenging phase diagram to the complex H-bond network, which has an important role not only in reactions in solution, but also in water ions dynamics. This is mainly due to the difficulty of modeling the delicate interplay among strong covalent bonds, weak van der Waals interactions and the H-bonds with a wide range of intensity. For this reason water is the perfect target of AIMD, particularly in the CPMD and BOMD formalisms [11, 42–45]. Within this framwork proton transfer in aqueous solution has been extensively simulated [43, 46–50].

Diagonal BO, which is just the adiabatic approximation with the diagonal contribution from the coupling matrix, has been limited to geometry optimization [51] and the study of vibrational states [52]. We note in passing that the field of nonadiabatic molecular dynamics, which extends beyond the the framework described in Section 1.1, is an active field of research, with intriguing applications in water ionization [53–55], eventually concerning proton transfer within ionized water systems [56–58]. This approach enables a more accurate comparison with experimental studies of water photodissociation in small, nevertheless neutral, clusters. As mentioned in the Introduction, in this thesis we are not dealing with such systems and dynamics. The protonated water clusters taken into account in this work have a total charge of $+e$ and do not interact with an external perturbation, hence ground state BOMD will be our workhorse. In particular the ground state PES is the mathematical object that we will first compute with the electronic structure calculations explained in Chap. 2, in order to use it in the nuclei classical and quantum dynamics (Chap. 3). Then, in the second part of this thesis, we will fit the PES with methods reported in Chap. 6.

Chapter 2

Electronic structure methods

Electronic structure methods aim to solve the electronic Schrödinger equation, in order to find the energy eigenvalues and differentiate them with respect to the nuclear position to obtain the forces necessary for the dynamics. We rewrite the time-independent Schrödinger equation with the clamped-nuclei Hamiltonian (Eq. 1.9), by using the explicit form for the kinetic and potential operators. We then have:

$$\left[\frac{\hbar}{2m} \sum_{i=1}^{N} \nabla_i^2 + \frac{1}{2} \sum_{i,j}^{N} \frac{e^2}{|\mathbf{r}_i - \mathbf{r}_j|} - \frac{1}{2} \sum_{i,a}^{NM} \frac{Z_a e^2}{|\mathbf{r}_i - \mathbf{q}_a|} \right] \Phi(\mathbf{r};\mathbf{q}) = E(\mathbf{q}) \Phi(\mathbf{r};\mathbf{q}), \tag{2.1}$$

where we ignored the nuclear interaction term $\hat{V}_{\text{nn}}$, defined in Eq. (1.3), as it acts on the nuclear coordinates only, and it sums just as a classical additive constant to the electronic energy. For the sake of readability, we will often drop the dependence of the wavefunction on the electrons degrees of freedom, $\mathbf{r}$, as well as its parametric dependence on the nuclear coordinates, $\mathbf{q}$. In the next Section 2.1 we will see how to get the forces. In Section 2.2 we will briefly overview what are the most common techniques in quantum chemistry, in order to motivate our choice of an alternative method in this first part of thesis, namely quantum Monte Carlo, which is the topic of Section 2.3.

2.1 Forces via the Hellmann-Feynman Theorem

The Hellmann-Feynman theorem provides a simple way of computing forces, once the wavefunction is found. It applies to any derivative of the expectation value of the Hamiltonian with respect to any of its parameters. In our case we are interested in the nuclear position parameter, $\mathbf{q}$:

$$-\mathbf{f} = \nabla_{\mathbf{q}} \langle \Phi | \hat{H} | \Phi \rangle = \langle \partial_{\mathbf{q}} \Phi | \hat{H} | \Phi \rangle + \langle \Phi | \partial_{\mathbf{q}} \hat{H} | \Phi \rangle + \langle \Phi | \hat{H} | \partial_{\mathbf{q}} \Phi \rangle. \tag{2.2}$$

By grouping together the derivatives on the bra and the ket, we obtain:

$$\langle \partial_{\mathbf{q}} \Phi | \hat{H} | \Phi \rangle + \langle \Phi | \hat{H} | \partial_{\mathbf{q}} \Phi \rangle = \partial_{\mathbf{q}} \langle \Phi | \Phi \rangle = 0, \tag{2.3}$$

since the wavefunction is normalized . This leaves us with only the middle term:

$$\frac{dE(\mathbf{q})}{d\mathbf{q}} = \left\langle \Phi(\mathbf{q}) \left| \frac{\partial \hat{H}(\mathbf{q})}{\partial \mathbf{q}} \right| \Phi(\mathbf{q}) \right\rangle, \tag{2.4}$$

2.2 Deterministic quantum chemistry methods

The first distinction to make is between wave function-based methods and electronic density-based methods. The former attempt to solve the eigenvalue problem by approximating the electronic wavefunction, Φ, while the latter calculate the energy as a functional of the electronic density, which has the advantage of depending on just three spatial coordinates. In this Section we will review both approaches, some of which will be employed in the second Part of this thesis. We begin with a central principle which is common to both methods.

2.2.1 The variational principle

Variational principles are omnipresent in physics. In particular, the formulation used to solve eigenvalues problems in the context of wave mechanics, due to Rayleigh [59] and Ritz [60], has found broad application later also in quantum chemistry.

The Rayleigh-Ritz variational principle in quantum mechanics states that given any normalized state $|\Phi\rangle$ of a many-body system belonging to the Hilbert space where a given Hamiltonian $\hat{H}_\mathrm{e}$ acts, one always has that

$$\langle \Phi | \hat{H} | \Phi \rangle \geq E_0, \tag{2.5}$$

where E_0 is the ground state energy and the equality holds only in the case $|\Phi\rangle = |\Phi_0\rangle$, with $|\Phi_0\rangle$ defined as the ground-state. This principle provides a way to find the ground-state wavefunction, that is by energy minimization:

$$E_0 = \min_{\Phi} \frac{\langle \Phi | H | \Phi \rangle}{\langle \Phi | \Phi \rangle}. \tag{2.6}$$

Usually the wavefunction depends on one or more parameters with respect to which the energy is minimized. Among these methods, those based on independent-particle approximation, namely the Hartree and the Hartree-Fock methods, are the starting point of many other more advanced techniques.

2.2.2 Independent-particle approaches

In this brief overview we primarily follow the classic textbook by Szabo and Ostlund [61]. We consider N electrons described by the variables $\mathbf{x}_i = (\mathbf{r}_i, s_i)$, where the $\mathbf{r}$ represents the position and s_i the spin of the electron.

In the absence of spin-orbit interaction, we can use a set K spatial *molecular orbitals* (MOs) $\{\psi_i^{\mathrm{MO}} | i = 1, 2, \cdots, N/2\}$ and two orthonormal spin functions $\alpha(\sigma)$ and $\beta(\sigma)$ to define the set of N

spin orbitals as a product between the spatial and spin functions:

$$\begin{cases} \chi_{2i-1}(\mathbf{x}) & = \psi_i^{\mathrm{MO}}(\mathbf{r})\alpha(\sigma) \\ \chi_{2i}(\mathbf{x}) & = \psi_i^{\mathrm{MO}}(\mathbf{r})\beta(\sigma) \end{cases}, \tag{2.7}$$

which will inherit the orthonormality from the MOs. The latter are typically constructed as linear combination of atomic orbitals (LCAO), which depend on the vector distance from the nucleus $\mathbf{q}_a$ to the electron $\mathbf{r}_i$, that is, $\mathbf{r}_a = \mathbf{r}_i - \mathbf{q}_a$. Approximations of atomic orbitals (AOs) basis functions, also called *primitives*, are usually expressed as the product of an angular component that depends on the direction $\hat{\mathbf{r}}_a$, such as spherical harmonics Y_l^m, and a radial component that depends just on the distance $r_a = |\mathbf{r}_a|$. For example, in the case of Slater-type orbitals (STOs) [62] and Gaussian-type orbitals (GTOs) [63], the forms are given by the following equations: which are usually constructed as linear combination of atomic orbitals (LCAO).

$$\psi_{a,lmn}^{\mathrm{STO}}(\mathbf{r}_a) \propto r_a^{n-1} e^{-\zeta r_a} Y_l^m(\hat{\mathbf{r}}_a), \tag{2.8}$$

$$\psi_{a,lmn}^{\mathrm{GTO}}(\mathbf{r}_a) \propto r_a^{l} e^{-\zeta r_a^2} Y_l^m(\hat{\mathbf{r}}_a), \tag{2.9}$$

respectively. The principal quantum number n limits the range of the angular momentum quantum numbers, l and m, with $l \in [0, n-l]$ and $m \in [-l, +l]$, used to define the spherical harmonics Y_l^m. For GTOs orbitals, in some cases, the priority is given to the choice of l, with $n \in [1, n_l]$ designating the number of Gaussians for each angular momentum shell.

Once a local basis of N_b AOs is defined, $\{\psi_k^{\mathrm{AO}}\}$, the MOs of a system of M atoms are defined as

$$\psi_i^{\mathrm{MO}}(\mathbf{r}) = \sum_{j=1}^{N_b \times M} \mu_{ij} \psi_j^{\mathrm{AO}}(\mathbf{r}), \tag{2.10}$$

where the AOs are indexed according to local basis and to the specific atom they belong to.

In the Hartree method [64] the many-body wave function is expressed as a simple product of single-particle wave functions, defined as the spin orbitals in Eq. (2.7):

$$\Phi(\mathbf{r}) = \chi_1(\mathbf{x}_1)\chi_2(\mathbf{x}_2)\cdots\chi_N(\mathbf{x}_N). \tag{2.11}$$

If we plug this wavefunction into equation (2.1), we obtain the energy:

$$\begin{aligned} E_{\mathrm{Hartree}} = & \sum_i^N \left[-\frac{1}{2} \int \mathrm{d}\mathbf{r} \psi_i^*(\mathbf{r}) \nabla_i^2 \psi_i(\mathbf{r}) \right] \\ & - \sum_i^N \left[\sum_a^M Z_a \int \mathrm{d}\mathbf{r} \psi_i^*(\mathbf{r}) \frac{1}{|\mathbf{r} - \mathbf{q}_a|} \psi_i(\mathbf{r}) \right] \\ & + \sum_i^N \left[\frac{1}{2} \sum_{i \neq j}^N \int\int \mathrm{d}\mathbf{r}\, \mathrm{d}\mathbf{r}' \psi_i^*(\mathbf{r}) \psi_j^*(\mathbf{r}') \frac{1}{|\mathbf{r} - \mathbf{r}'|} \psi_i(\mathbf{r}) \psi_j(\mathbf{r}') \right], \end{aligned} \tag{2.12}$$

where the three lines correspond to the contributions from the electrons kinetic energy, the attractive ion-electron interaction and the Coulombic repulsion between electrons. The latter, also

called *Hartree direct term*, E_{H}, sums over all the possible products between the square moduli of two wavefunctions, which gives the joint probability of two electrons being in the same position. From these results, where the spins degrees of freedom are integrated out, we see that the problem with the Hartree method is that the wavefunction does not respect the Pauli principle.

This issue is addressed by the Hartree-Fock (HF) approximation [65], where the wavefunction is represented by a single Slater determinant [66], which incorporates the anti-symmetry required by the Pauli principle:

$$\Phi(\mathbf{r}) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \chi_1(\mathbf{x}_1) & \chi_2(\mathbf{x}_1) & \cdots & \chi_N(\mathbf{x}_1) \\ \chi_1(\mathbf{x}_2) & \chi_2(\mathbf{x}_2) & \cdots & \chi_N(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(\mathbf{x}_N) & \chi_2(\mathbf{x}_N) & \cdots & \chi_N(\mathbf{x}_N) \end{vmatrix}. \tag{2.13}$$

The resulting energy in the HF method is the same as in the Hartree method, with the addition of an an *exchange* term E_{X},

$$E_{\mathrm{HF}} = E_{\mathrm{Hartree}} + E_{\mathrm{X}}, \tag{2.14}$$

with

$$E_{\mathrm{X}} = -\frac{1}{2} \sum_{i \neq j}^{N} \int \mathrm{d}\mathbf{r}\, \mathrm{d}\mathbf{r}' \phi_i^*(\mathbf{r}) \mathbf{r} \phi_j^*(\mathbf{r}') \frac{1}{|\mathbf{r} - \mathbf{r}'|} \phi_i(\mathbf{r}') \phi_j(\mathbf{r}). \tag{2.15}$$

Notice the different order of the integration variables compared to the last term in Eq. (2.12)). In this case, we are not dealing with the product of two electronic densities to account for their Coulombic interaction. Instead, thanks to the Slater determinant, the HF method is able to consider a purely quantum contribution to the energy, specifically the one due to the motion of two electrons with parallel spins.

However, the methods discussed so far are considered *uncorrelated methods*, because they do not account for the full *correlation energy* E_C, formally defined as

$$E_C := E_0 - E_{\mathrm{HF}}^0, \tag{2.16}$$

where E_0 the true energy of the ground state, and E_{HF}^0 the Hartree-Fock energy in the infinite-basis limit, meaning that the Slater derminant is composed using linear combination of an infinite number of MOs.

The correlation energy has two components:

- *Static correlation*, which arrises when the correct electronic structure requires multiple configurations, thus multiple determinants, to adequately describe the system. This is particularly important for bond dissociation.

- *Dynamic correlation*, which is related to the instantaneous interaction between electrons as they move. Unlike static correlation, it does not have a multi-configurational character, but it is necessary to correctly describe the electron-electron repulsion, especially in the case of antiparallel spin.

The purpose of advanced computational methods is to recover this fundamental contribution to the energy, in order to correctly describe various phenomena that occur at scale of chemical accuracy.

2.2.3 Beyond Hartree-Fock: correlation energy in wavefunction methods

Post-Hartree-Fock methods constitute a vast group of deterministic computational chemistry techniques aimed at incorporating correlation energy to some extent. Although these methods are not directly employed in this thesis, except for a fitted potential described in Section 5.2 that is based on them, they are worth mentioning to justify our methodology.

1. The most straightforward way to improve upon HF is to include more determinants, where one or more single-particle wavefunction are substituted by excited states. Two of these methods are the Configuration Interaction (CI), where a linear combination of multiple Slater determinants is optimized to recover E_C, and Multi-Configuration Self-Consistent Field (MCSCF), in which also the molecular orbitals are optimized. Being multi-configurational by definition, these methods effectively capture the static correlation, but exhibit a slow convergence when accounting for dynamic correlation, requiring a large number of Slater determinants in the expansion. Full-CI represents the theoretical limit of considering an infinite sum of Slater determinants, and is rarely applied beyond diatomic and triatomic sytems. Like HF, these methods still rely on the variational principle and in numerical analysis would be collectively indicated as Galerkin methods, because the solution of the differential equation is approximated by projecting it onto a finite-dimensional subspace spanned by the finite basis.

2. Møller-Plesset perturbation theory (MPPT) [67] is based on a perturbative expansion of the wavefunction around the HF solution Φ_0^{HF}. The second-order perturbation approximation (MP2) is the most common level of approximation, using the lowest non-vanishing correction term. Its computational costs is the lowest among the post-HF methods, estimated at $\mathcal{O}(N^5)$, with N the number of electrons.

3. Coupled Cluster (CC) theory [68–70], considered the "golden standard" of quantum chemistry computational methods, is based on applying the exponential of the excitation operator $\hat{T}$ to the HF wave function, $\Phi_{\mathrm{CC}} = e^{\hat{T}}\Phi_0^{\mathrm{HF}}$, allowing contribution from singly excited, doubly excitated, and higher-order determinants. The most common variants are those that account for single and double excitation (CCSD), with the option to include triple excitations computed perturbatively in CCSD(T). In the latter case, the computational costs is $\mathcal{O}(N^7)$ [71].

2.2.4 Density functional theory

The electronic density is defined as

$$\rho(\mathbf{r}) = N \int d\mathbf{r}_1 \cdots \mathbf{r}_N \left(\sum_{i=1}^{N} \delta(\mathbf{r} - \mathbf{r}_i) \right) |\Phi(\mathbf{r}_1, \cdots, \mathbf{r}_N)|^2 \tag{2.17}$$

Density Functional Theory (DFT) is the most widely used electronic structure method in ab initio chemistry. and it is founded on two important theorems by Hohenberg and Kohn [72]. A comprehensive and detailed treatment of DFT can be found in [73], from which we derive the basic concepts.

First HK theorem

For an interacting system of electrons subjected to an external potential $V_{ext}(\mathbf{r})$, the latter is fully and uniquely determined, up to an additive constant, by the electronic ground state density $\rho_0(\mathbf{r})$.

As a Corollary, determining the external potential also fully determines the Hamiltonian, and thus all the many-body wavefunctions for all the states, from the ground state to the excited ones. Therefore, all properties of the systems are completely determined by the ground-state density $\rho_0(\mathbf{r})$

Second HK theorem

For a given external potential $\hat{V}_{ext}$ the energy of the ground state is given by the global minimum of the energy functional, defined as

$$E\left[\rho_0(\mathbf{r})\right] = \underbrace{\hat{T}_{\mathrm{e}}\left[\rho_0(\mathbf{r})\right] + \hat{V}_{\mathrm{ee}}\left[\rho_0(\mathbf{r})\right]}_{F_{\mathrm{HK}}[\rho_0]} + \hat{V}_{\mathrm{ext}}\left[\rho_0(\mathbf{r})\right]. \tag{2.18}$$

where we have defined the universal *functional $F_{HK}[\rho_0]$, which is the same for electron systems, independent of the external potential.*

Thus, knowledge of the functional is sufficient to determine the ground state and the electronic density of the system. In our case, the external potential is that due to the presence of the nuclei, i.e. the electron-nuclear interaction, $\hat{V}_{\mathrm{ext}} = \hat{V}_{\mathrm{ne}}$.

Kohn-Sham equations

Minimizing the energy functional in Eq. 2.18 is non-trivial, due to the presence of many-body terms in the electron-electron interaction V_{ee}. Kohn and Sham proposed an elegant solution to this problem [74], which has become the standard tool in DFT.

Their approach maps the many-body problem onto a single-particle problem characterized by the same electronic density. According to the HK theorems, this *auxiliary system* will have the same ground state energy of the real system of interest. The ground state energy of the

auxiliary system, E_s, is described by a functional similar to the one we already encountered:

$$E_s\left[\rho(\mathbf{r})\right] = T_s\left[\rho(\mathbf{r})\right] + E_{\mathrm{H}}\left[\rho(\mathbf{r})\right] + E_{\mathrm{XC}}\left[\rho(\mathbf{r})\right] + E_{\mathrm{ext}}\left[\rho(\mathbf{r})\right], \tag{2.19}$$

where $E_{\mathrm{H}}\left[\rho(\mathbf{r})\right]$ is the Hartree functional, analogous to the Hartree direct term of Eq. 2.12, $E_{\mathrm{XC}}\left[\rho(\mathbf{r})\right]$ is the *exchange-correlation* term, which account not only for the exchange contribution, as in Eq. 2.15, but also for additional correlation effects. Finally, $T_s\left[\rho(\mathbf{r})\right]$ and $E_{\mathrm{ext}}\left[\rho(\mathbf{r})\right]$ represent the usual kinetic and external potential terms, respectively.

The exchange-correlation functional is defined as

$$E_{\mathrm{XC}}\left[\rho(\mathbf{r})\right] = F_{\mathrm{HK}}\left[\rho(\mathbf{r})\right] - \left(T_s\left[\rho(\mathbf{r})\right] + E_{\mathrm{H}}\left[\rho(\mathbf{r})\right]\right). \tag{2.20}$$

If we explicit the universal functional F as we defined it in the Second HK theorem in Eq. (2.18), we get:

$$E_{\mathrm{XC}}\left[\rho(\mathbf{r})\right] = T\left[\rho(\mathbf{r})\right] - T_s\left[\rho(\mathbf{r})\right] + V_{\mathrm{ee}}\left[\rho(\mathbf{r})\right] + E_{\mathrm{H}}\left[\rho(\mathbf{r})\right]), \tag{2.21}$$

we notice that the XC functional accounts for everything that cannot be described by the Hartree and HF methods.

The auxiliary independent-particle system automatically defines the single-particle auxiliary Hamiltonian,

$$\hat{H}_s = -\frac{1}{2}\nabla + V_s(\mathbf{r},\sigma), \tag{2.22}$$

which consists of the kinetic energy operator and an effective local potential that depends on electron position $\mathbf{r}$ and spin σ. Since this is an independent particle Hamiltonian, the ground state solution is determined by the electrons occupying first N eigenfunctions $\chi_i(\mathbf{r},\sigma)$ of $\hat{H}_s$ with the lowest eigenvalues:

$$\hat{H}_s\chi_i(\mathbf{r},\sigma) = \epsilon_i\chi_i(\mathbf{r},\sigma). \tag{2.23}$$

Given the eigenfunctions, the definition of the density is straighforward:

$$\rho(\mathbf{r}) = \sum_i^N |\chi_i(\mathbf{r},\sigma)|^2. \tag{2.24}$$

The idea behind the Kohn-Sham variational approach is to minimize the energy functional of the auxiliary system with respect to the density ρ, defined as in Eq. 2.24:

$$\frac{\delta E_s}{\delta\chi_i(\mathbf{r},\sigma)} = \frac{\delta T_s}{\delta\chi_i(\mathbf{r},\sigma)} + \underbrace{\left[\frac{\delta E_{\mathrm{ext}}}{\delta\rho(\mathbf{r},\sigma)} + \frac{\delta E_{\mathrm{Hartree}}}{\delta\rho(\mathbf{r},\sigma)} + \frac{\delta E_{\mathrm{XC}}}{\delta\rho(\mathbf{r},\sigma)}\right]}_{V_{\mathrm{ext}}(\mathbf{r})+V_{\mathrm{Hartree}}(\mathbf{r})+V_{\mathrm{XC}}(\mathbf{r},\sigma)=V_s(\mathbf{r},\sigma)} \frac{\delta\rho(\mathbf{r},\sigma)}{\delta\chi_i(\mathbf{r},\sigma)} = 0. \tag{2.25}$$

Being the wavefunctions subjected to the orthonormalization constraints, this minimization problem is analogous to the Rayleigh-Ritz variational approach for wavefunctions. In Eq. 2.25 we have grouped some functional derivatives into the effective potential that appears in 2.22.

$$V_s(\mathbf{r},\sigma) = V_{\mathrm{ext}}(\mathbf{r}) + V_{\mathrm{Hartree}}(\mathbf{r}) + V_{\mathrm{XC}}(\mathbf{r},\sigma) \tag{2.26}$$

The KS equations 2.22, 2.26 and 2.25 are solved self-consistently: starting from an effective potential V_s, the KS Hamiltonian is solved to obtain the electronic density, which is then used to compute a new effective potential. This loop is iterated until the difference in electronic density difference, between two iterations becomes smaller than a user-defined threshold.

Choosing the right exchange-correlation functionals for water

We conclude this overview on DFT with one of its most critical aspects: choosing the right exchange-correlation functional, particularly for water simulations. The XC functional E_{XC} is the key approximation of DFT, and selecting the appropriate one significantly affects the accuracy of the results. However, more precise functionals often come with increased computational cost.

Common choices include the Local Density Approximation (LDA), where the E_{XC} is given by the sum of the Slater exchange energy of the HF formula (Eq. 2.15), while the local correlation energy is fitted to the one of the homogeneous electrons gas, determined through accurate quantum Monte Carlo simulations at different values of density $\rho(\mathbf{r})$ [75, 76]; and the Generalized Gradient Approximation (GGA), which are semi-local, as it also accounts for inhomogeneity in the electron density via the gradient of the density, $\nabla\rho(\mathbf{r})$. Examples of GGA functionals include the Perdew-Burke-Ernzerhof (PBE) one [77], and the combination of the B88 exchange functional [78] and Lee-Yang-Parr [79] correlation functional (BLYP). Despite their widespread use, LDA performs poorly in water simulations as it overastimates the binding energy of the water clusters [80, 81]. This issue is due to spurious exchange attraction at large distances [82]. GGAs functionals also have drawbacks, particularly in over-structuration of bulk liquid water, which translates into a small diffusion constant, an overly large average number of H-bonds, and a liquid phase less dense than the ice [83, 84]. Although these overbinding [85] effects are more pronounced in the bulk than in water cluster, H-bonds play too crucial role in proton hopping to be poorly reproduced.

The fact that local and semi-local XC functionals perform better in gas-phase water cluster than in bulk water suggests the necessity of including many-body effects such as van der Waals interactions. This can be done in several ways, the most simple being adding an atom-atom attractive semiempirical pair potential having the London dispersion functional form, $-C_6/R^6$ [86]. Another approach consists in including in the XC functional a non-local correlation term E_C^{nl} that depends explicitly on the electron densities at spatially separated positions. These XC functionals, generally dubbed van der Waals Density Functionals (vdW-DF), are defined as follows:

$$E_{XC} = E_X^{GGA} + E_C^{LDA} + E_C^{nl}, \tag{2.27}$$

where the first term is a GGA exchange term [87], the second one is the Perdew-Wang local correlation (PW86) [88], and the last term is the non-local contribution generally defined as

$$E_C^{nl}[\rho] = \int d\mathbf{r}\, d\mathbf{r}' \rho(\mathbf{r})\phi(\mathbf{r},\mathbf{r}')\rho(\mathbf{r}') \tag{2.28}$$

where $\phi(\mathbf{r},\mathbf{r}')$ is a function of $1/|\mathbf{r}-\mathbf{r}'|$, $\rho(\mathbf{r})$, $\rho(\mathbf{r}')$, and their gradients. The specific form of $\phi(\mathbf{r},\mathbf{r}')$ defines the type of vdW functional. In the second Part of this thesis we will use the vdW-DF2 [89], which represents an improvement over the original vdW-DF [90], particularly for systems with shorter range dispersion forces.

Inclusion of vdW interaction reduces the gap between DFT and advanced quantum chemistry methods [91], with an improved reproduction of oxygen-oxygen radial distribution func-

tion with respect to GGA [92]. However vdW-DF2 functionals are not exempt from defects, as they tend to understructure liquid water [93].

2.3 Quantum Monte Carlo

Quantum Monte Carlo (QMC) is a family of stochastic integration algorithms aimed to solve various quantum problems. We refer the reader to [94, 95] for an introduction and to [96, 97] for review papers. In this context, we are particularly interested in the Variational Monte Carlo (VMC) variant, which is one of the earliest QMC methods [98, 99]. In VMC, a trial many-body wavefunction of the electrons, $\Phi_T(\mathbf{r}) = \Phi_T(\mathbf{r}_1, \cdots, \mathbf{r}_N)$, is first optimized using the variational theorem, as other methods discussed earlier, and then used to estimate the variational energy and other observables.

More precisely, the quantum expectation value of the Hamiltonian $\hat{H}_\mathrm{e}$ is computed with the trial wavefunction Φ_T according to

$$\frac{\langle \Phi_T | H_\mathrm{e} | \Phi_T \rangle}{\langle \Phi_T | \Phi_T \rangle} \equiv \left\langle \hat{H}_\mathrm{e} \right\rangle = \frac{\int \mathrm{d}\mathbf{r} \Phi_T^*(\mathbf{r}) \hat{H} \Phi_T(\mathbf{r})}{\int \mathrm{d}\mathbf{r} |\Phi_T(\mathbf{r})|^2} = E_\mathrm{VMC}, \tag{2.29}$$

where $\mathbf{r}$ is again understood as the vector of all electron coordinates, $(\mathbf{r}_1, \cdots, \mathbf{r}_N)$. By changing variable in the denominator, $\mathbf{r} \to \mathbf{r}'$, and by multiplying both numerator and denominator by $\Phi_T(\mathbf{r})$, we can rewrite the expression of the VMC energy as

$$E_\mathrm{VMC} = \int \mathrm{d}\mathbf{r} \frac{|\Phi_T(\mathbf{r})|^2}{\int \mathrm{d}\mathbf{r}' |\Phi_T(\mathbf{r}')|^2} \frac{\hat{H}\Phi_T(\mathbf{r})}{\Phi_T(\mathbf{r})} = \int \mathrm{d}\mathbf{r} \pi(\mathbf{r}) E_L(\mathbf{r}) = \langle E_L \rangle \geq E_0, \tag{2.30}$$

where we have defined the local energy

$$E_L = \frac{\hat{H}\Phi_T(\mathbf{r})}{\Phi_T(\mathbf{r})}, \tag{2.31}$$

which is sampled according to the following probability distribution

$$\pi(\mathbf{r}) = \frac{|\Phi_T(\mathbf{r})|^2}{\int \mathrm{d}\mathbf{r}' |\Phi_T(\mathbf{r}')|^2}. \tag{2.32}$$

Therefore, VMC is an *importance sampling* technique, as the electronic configurations are not sampled uniformly, but rather according to the amplitude of the trial wavefunction, $|\Phi|^2$.

In practice the probabily density $\pi(\mathbf{r})$ is sampled using standard Monte Carlo Markov Chains (MCMC) methods, such as the Metropolis-Hastings algorithm [100–102]. MCMC is a random walk that sample the unknown probability distributions defined on a configuration space by jumping from one configuration to another depending only on the current one. This method is particularly suited for solving integrals in high-dimensional spaces, as it is the case for the 3N-dimensional configuration space of the electronic degrees of freedom.

Given an observable $\mathcal{O}$, the sample mean $\overline{O}$ of N_gen configurations is an *unbiased* estimator of the population mean $\langle \mathcal{O} \rangle$,

$$\langle \mathcal{O} \rangle \approx \overline{O} = \frac{1}{N_\mathrm{gen}} \sum_{i=1}^{N_\mathrm{gen}} \hat{O}_L(\mathbf{r}_i), \tag{2.33}$$

where $\hat{O}_L = \frac{\hat{O}\Phi_T}{\Phi_T}$ is the local operator corresponding to the observable of interest, averaged of the $\{\mathbf{r}_i\}$ configurations distributed according to $\phi(\mathbf{r})$. For example, the energy E_{VMC} is estimated through the local one E_L:

$$E_{\text{VMC}} = \langle E \rangle \approx \overline{E} = \frac{1}{N_{\text{gen}}} \sum_{i=1}^{N_{\text{gen}}} E_L(\mathbf{r}_i). \tag{2.34}$$

By virtue of the central limit theorem (CLT), we know that if the random variables $E_L(\mathbf{r}_i)$ are *independent and identically distributed* (iid), and if $\pi(\mathbf{r})$ has a finite expected value $\mathbb{E}[E_L]$ and finite variance $\text{var}[E_L] = \mathbb{E}\left[(E_L - E_{\text{VMC}})^2\right]$, then in the limit of $N_{\text{gen}} \to \infty$ the sample mean $\overline{E}_L$ converges to a Gaussian distribution with the following expected value and variance:

$$\mathbb{E}\left[\overline{E}\right] = \mathbb{E}[E_L] = E_{\text{VMC}}, \tag{2.35}$$

$$\text{var}[\overline{E}] = \frac{\text{var}[E_L]}{N_{\text{gen}}}. \tag{2.36}$$

From the Eq. (2.36) we can see the strength of MC methods over deterministic ones: the intrinsic statistical error,

$$\sigma[\overline{E}] = \sqrt{\frac{\text{var}[E_L]}{N_{\text{gen}}}}, \tag{2.37}$$

depends only on the number of MC iterations, and not on the dimensionality of the integral.

Moreover, there are two interesting properties in the specific case of quantum Monte Carlo. The *zero variance* property states that, in the limit the wavefunction Φ approaching the exact eigenfunction of $\hat{H}_e$, the local energy E_L will also approach the exact value and becomes independent of $\mathbf{r}$, with the statistical uncertainty of $\overline{E}_L$ vanishing. The *zero-bias* property implies that the systematic error of the variational energy with respect to the exact energy E_0 vanishes in the same limit.

The caveat is that MCMC provides correlated samples of the local operators, meaning that two electronic configuration sampled at two step whose distance is smaller than a certain *autocorrelation time* will not be independent. A simple and elegant statistical method to take into account the correlation between samples is the *block averaging* [103] technique. We divide the whole sampling in N_B blocks, each containing N_s samples. The average within the block will be simply:

$$\overline{O}_B = \frac{1}{N_s} \sum_{i=1}^{N_s} O_i, \tag{2.38}$$

while the total average is the average over the blocks:

$$\overline{O} = \frac{1}{N_B} \sum_{b=1}^{N_B} \overline{O}_b. \tag{2.39}$$

where the subscript b runs from 1 to the last block N_B. If the blocks size N_s is larger than the correlation time, it is safe to compute the variance of the sample mean as

$$\text{var}\left[\overline{O}\right] = \frac{\text{var}\left[\overline{O}_b\right]}{N_B} \tag{2.40}$$

which in practice translates to the following standard formula:

$$\mathrm{var}\left[\overline{O}\right] \approx \frac{1}{N_B - 1}\left[\frac{1}{N_B}\sum_{b=1}^{N_b}\overline{O}_b^2 - \left(\frac{1}{N_B}\sum_{b=1}^{N_b}\overline{O}_b\right)^2\right]. \tag{2.41}$$

The appropriate block size N_s can be determined heuristically by running the average for several size values and identifying the value of N_s the variance no longer increase. Alternatively, one can estimate the autocorrelation time by considering the variance over the entire set. Calling the local value of an operator $O_i = O(\mathbf{r}_i)$ for short, we have

$$\mathrm{var}\left[O\right] = \frac{1}{N_{\mathrm{gen}}^2}\sum_{i,j}^{N_{\mathrm{gen}}}\mathrm{cov}\left[O_i, O_j\right], \tag{2.42}$$

where we used the *normalized time autocorrelation function*

$$\mathrm{cov}\left[O_i, O_j\right] = \left\langle \left(O_i - \langle O\rangle\right)\left(O_j - \langle O\rangle\right)\right\rangle. \tag{2.43}$$

The formula can be approximated by considering absolute "time" distance between two samples:

$$\mathrm{var}\left[O\right] \approx \frac{1}{N_{\mathrm{gen}}^2}\sum_{i}^{N_{\mathrm{gen}}}\sum_{t=-\infty}^{+\infty} c(|t|) = \frac{\tau}{N_{\mathrm{gen}}}\sigma^2(O) \tag{2.44}$$

with the autocorrelation function redefined as:

$$c(t) = \left\langle O_s O_{s+t}\right\rangle - \left\langle O\right\rangle^2 \tag{2.45}$$

and

$$\tau = 1 + 2\sum_{t=1}^{\infty}\frac{c(t)}{c(0)} \tag{2.46}$$

is the *autocorrelation time* which give us an estimation of the number of effectively independent points in the whole sampled set.

2.3.1 Forces in quantum Monte Carlo

The evaluation of derivatives in quantum Monte Carlo is of paramount importance not only if one is interested in the forces for the dynamics, but also for variational energy minimization, which is usually done iteratively by using gradients, as we will see in Section 2.3.2. Unfortunately, computing forces in QMC is not as straighforward as in other methods like DFT. Following the same procedure as in 2.1, if we compute

$$\mathbf{f} = -\nabla_{\mathbf{q}} E_{\mathrm{VMC}}[\Phi], \tag{2.47}$$

we must account for all the dependencies of the energy functional with respect to nuclear positions. In addition to dependence on $\mathbf{q}$ through the Hamiltonian, E_{VMC} has an explicit dependence through Φ, if the latter is defined with localized basis set, as is often the case, and an

implicit dependence through the variational parameters $\boldsymbol{\lambda}$, which are optimized for a given $\mathbf{q}$. By substituting Eq. 2.30 in Eq. 2.47, we can express the force as the sum of three contribution:

$$\mathbf{f} = \mathbf{f}^{\text{Hel-Fey}} + \mathbf{f}^{\text{Pulay}} + \mathbf{f}^{\boldsymbol{\lambda}}, \tag{2.48}$$

where

$$\mathbf{f}^{\text{Hel-Fey}} = -\left\langle \nabla_{\mathbf{q}} E_L \right\rangle_{\Phi_T} \tag{2.49a}$$

$$\mathbf{f}^{\text{Pulay}} = -2\left\langle (E_L - E_{\text{VMC}}) \left[\nabla_{\mathbf{q}} \log \Phi\right]\right\rangle_{\Phi_T} \tag{2.49b}$$

$$\mathbf{f}^{\boldsymbol{\lambda}} = -\nabla_{\boldsymbol{\lambda}} E_{\text{VMC}} \cdot \nabla_{\mathbf{q}} \boldsymbol{\lambda}. \tag{2.49c}$$

The first term is the usual Hellman-Feynman contribution, the second is the Pulay term and the last one contains the dependence on the variational parameter, and it is the most complicated to compute.

Fortunately, when the true energy minimum and the true ground state are reached, the $\mathbf{f}^{\boldsymbol{\lambda}}$ is zero by definition: $\frac{\partial E_{\text{VMC}}}{\partial \boldsymbol{\lambda}} = 0$.

In the same way, when the wavefunction approaches an eigenstate of $\hat{H}_{\text{e}}$, the Pulay term vanishes, leaving only the Hellman-Feynman contribution. However, in practice, the wavefunction is never an exact eigenstate of $\hat{H}_{\text{e}}$, and the Pulay stress poses a problem even in deterministic quantum chemistry methods, because the wavefunction is always approximated using a finite basis set.

Additionally, as with all observables in QMC, forces are computed as averages, which must have a finite variance. A naive application of finite difference derivatives, with the finite step Δ approaching zero, will end up in a diverging error on the forces, as the QMC energy difference error remains constant, while $\Delta \to 0$. This problem has been addressed using correlated sampling (CS) in VMC [104, 105] and DMC [106], and by Space-Warp Coordinate Transformation (SWCT) [105], which provides an estimator of the force with zero variance. With SWCT, the electronic coordinates $\mathbf{r}$ follow the nuclear ones $\mathbf{q}_a$ when these are displaced, mimicking the displacement of the charge around the nucleus. SWCT has been generalized to infinitesimal ion displacements via algorithmic differentiation (AD) [107], which made the computational cost of QMC forces only four times more expensive than the energy point calculation. Furthermore, SWCT has recently been thoroughly tested [108] and refined in the VMC case to provide very accurate forces for machine learning applications [109].

The issue of infinite variance is not limited only to the numerical approximations of the derivatives, but also affects the analytical differentiation. Indeed, the $\mathbf{f}^{\text{Hel-Fey}}$ term may diverge as electron-ion distance approaches zero, and the $\mathbf{f}^{\text{Pulay}}$ term diverges near the nodal surface, where $\Phi_T(\mathbf{r}) = 0$. Several variance reduction methods has been proposed to tackle this issue, specifically for the Hellman-Feynamn term [110, 111], or for both $\mathbf{f}^{\text{Hel-Fey}}$ and $\mathbf{f}^{\text{Pulay}}$ in the periodic boundary case [112] and the open one [113] as well.

2.3.2 Wavefunction optimization

The main difficulty in applying the variational principle for wavefunction optimization arises from the the fact that the target function, the energy, is known only statistically. Historically the problem has been tackled first by running several independent energy runs [98, 114], but this was limited in the number of variational parameters and by high computational cost.

A widely used iterative method for high-dimensional optimization is the steepest descent algorithm [1], which exploits the derivative information to drive the parameters towards the energy minimum. In our specific case, this would mean to use the f_k, i.e. the energy derivative with respect to the parameter λ_k, to update the same parameter according to

$$\lambda_k' - \lambda_k = \delta\lambda_k = -\Delta\frac{\partial E}{\partial \lambda_k} = \Delta f_k \tag{2.50}$$

which is equivalent to minimize the following cost function

$$\arg\min_{\boldsymbol{\lambda}} \left[E + \sum_k \left(-\delta\lambda_k f_k + \frac{1}{2\Delta}\delta\lambda_k^2 \right) \right]. \tag{2.51}$$

The issue with the steepest descent approach is that it assumes that all parameters are affected by thes same curvature, but often some parameters are more difficult to optimize.

The solution is to take into account the geometry of the parameter space by using an appropriate metric, such as the Fisher information matrix, to evaluate the local curvature and compute natural gradients [115]. This technique was introduced as the stochastic reconfiguration (SR) algorithm by Sorella in the context of Green function Monte Carlo [116], and was later extended to VMC [117, 118]. This method leverages the direct knowledge of the trial QMC wavefunction, particularly concerning the Hilbert space topology in which it is defined, to achieve rapid convergence. Here, we briefly describe the main ideas behind it [2].

Consider the variational parameter as a single vector of length p:

$$\boldsymbol{\lambda} = \left(\lambda_1, \lambda_2, \cdots, \lambda_p\right). \tag{2.52}$$

We can define the logarithm derivative operator O_k as:

$$\hat{O}_k(\mathbf{r}) = \frac{\partial}{\partial \lambda_k} \log \Phi_{\boldsymbol{\lambda}}(\mathbf{r}) = \frac{\partial_{\lambda_k}\Phi_{\boldsymbol{\lambda}}(\mathbf{r})}{\Phi_{\boldsymbol{\lambda}}(\mathbf{r})}, \tag{2.53}$$

for $\Phi_{\boldsymbol{\lambda}} \neq 0$.

We can express the variational wavefunction as

$$|\Phi_{\boldsymbol{\lambda}+\delta\boldsymbol{\lambda}}\rangle = |\Phi_{\boldsymbol{\lambda}}\rangle + \sum_k \delta\lambda_k \frac{\partial |\Phi_{\boldsymbol{\lambda}}\rangle}{\partial \lambda_k} + o(\delta\boldsymbol{\lambda}^2) = \left(1 + \sum_k \delta\lambda_k O_k\right)|\Phi_{\boldsymbol{\lambda}}\rangle + o(\delta\boldsymbol{\lambda}^2) \tag{2.54}$$

[1] A stochastic variant of this algorithm, the stochastic gradient descent, is presented in Chapter 6, devoted to machine learning, including the optimization of neural networks.

[2] The presentation here differs from the original as it emphasizes the analogies and the differences with the steepest descent method. This derivation comes from the Lecture notes of Michele Casula's course at the *TREX School on QMC with TurboRVB* organized by TREX and SISSA in July 2023.

$$
\begin{aligned}
|\Phi_{\boldsymbol{\lambda}+\delta\boldsymbol{\lambda}}\rangle &= |\Phi_{\boldsymbol{\lambda}}\rangle + \sum_k \delta\lambda_k \frac{\partial\,|\Phi_{\boldsymbol{\lambda}}\rangle}{\partial\lambda_k} + o(\delta\boldsymbol{\lambda}^2) = \\
&= \left(1 + \sum_k \delta\lambda_k O_k\right)|\Phi_{\boldsymbol{\lambda}}\rangle + o(\delta\boldsymbol{\lambda}^2) =
\end{aligned}
\tag{2.55}
$$

We are interested in the normalized wavefunction,

$$
\left|\tilde{\Phi}_{\boldsymbol{\lambda}}\right\rangle = \frac{|\Phi_{\boldsymbol{\lambda}}\rangle}{\|\Phi_{\boldsymbol{\lambda}}\|} \qquad \text{where} \qquad \|\Phi_{\boldsymbol{\lambda}}\| = \sqrt{\langle\Phi_{\boldsymbol{\lambda}}|\Phi_{\boldsymbol{\lambda}}\rangle}, \tag{2.56}
$$

and in quantifying how much it changes:

$$
\left|\delta\tilde{\Phi}\right\rangle = \left|\tilde{\Phi}_{\boldsymbol{\lambda}+\delta\boldsymbol{\lambda}}\right\rangle - \left|\tilde{\Phi}_{\boldsymbol{\lambda}}\right\rangle. \tag{2.57}
$$

For this, we use the normed variation of the wavefunction ds^2 defined as

$$
ds^2 = \left\|\left|\tilde{\Phi}_{\boldsymbol{\lambda}+\delta\boldsymbol{\lambda}}\right\rangle - \left|\tilde{\Phi}_{\boldsymbol{\lambda}}\right\rangle\right\|^2 = \left\langle\delta\tilde{\Phi}|\delta\tilde{\Phi}\right\rangle. \tag{2.58}
$$

Inserting Eq. 2.57 into Eq. 2.58 we get:

$$
ds^2 = \sum_{kk'} \delta\lambda_k \delta\lambda_{k'} \left\langle\tilde{\Phi}_{\boldsymbol{\lambda}}\right| \left(O_k - \overline{O_k}\right)\left(O_{k'} - \overline{O_{k'}}\right)\left|\tilde{\Phi}_{\boldsymbol{\lambda}}\right\rangle = \sum_{kk'} \delta\lambda_k \delta\lambda_{k'} S_{kk'}, \tag{2.59}
$$

where we have defined the *stochastic reconfiguration matrix* from the covariance matrix of the logarithm derivative operator:

$$
S_{kk'} = \mathrm{cov}\left[O_k, O_{k'}\right] \tag{2.60}
$$

which is also known as Fisher information metric $F = 4S$ of the probability $p_{\boldsymbol{\lambda}}(x) \propto \Phi_{\boldsymbol{\lambda}}(\mathbf{x})^2$. Thus, instead of using the Euclidean metric, we can use the more appropriate Fisher information metric to define our cost function:

$$
\arg\min_{\delta\boldsymbol{\lambda}} \left[-\sum_k f_k \delta\lambda_k + \frac{ds^2}{2\Delta}\right], \tag{2.61}
$$

or, in matrix form:

$$
\arg\min_{\delta\boldsymbol{\lambda}} \left[\mathbf{f}\delta\boldsymbol{\lambda} + \frac{1}{2\Delta} S\right] \tag{2.62}
$$

from which we get the solution:

$$
\delta\boldsymbol{\lambda} = \Delta S^{-1}\mathbf{f} \tag{2.63}
$$

where $\mathbf{f}$ is the vector of energy derivatives. Notice that its expression is given by Eq. (2.49b), as it is equivalent to the energy derivatives with respect to the ionic positions, with the notable difference that the Helmann-Feynman contribution is zero because in this case only the wave function, and not the Hamiltonian, depends on the parameters $lambda_k$.

2.3.3 The wave function ansatz

As in Section 2.2.2, we describe N electrons using generalized coordinates, collectively indicated as $\mathbf{x} = \{\mathbf{x}_i\} = \{(\mathbf{r}_i, \sigma_i)\}_{i=1,\cdots,N}$, while $\mathbf{r}$ stands for all the space coordinates only. For convenience we restrict to the case of spin-unpolarized system, that is $N_\uparrow = N_\downarrow = N/2$, but the same approach has been applied also to the spin-polarized case [119, 120].

The wavefunction used in this work is the product of two contributions:

$$\Phi(\mathbf{x}) = \Phi_{\text{AS}}(\mathbf{x}) \times e^{J(\mathbf{x})} \tag{2.64}$$

where J is the *Jastrow factor*, a bosonic function of the electron degrees of freedom [121], while Φ_{AS} is an antisymmetric function, thus fermionic, and it is also referred to as *determinantal* part of the WF, because the easiest way to encode antisymmetry is through one ore more Slater determinant. Such a compact form with the Jastrow in exponential form make ensure a rapid convergence of the energy despite a large number of parameters $\boldsymbol{\lambda}$. In the following we describe the functional form of each factor.

Antisymmetrized geminal power

The antisymmetric part can be built in different ways, the most straightforward would be a single Slater determinant. In our case we consider a generalization of the Resonating Valence Bond (RVB) wavefunction, first proposed by Pauling in quantum chemistry [122] to describe aromatic molecules, and later reprised Anderson [123] in condensed matter to describe strongly correlated system. The RVB-WF describes a superposition of all possible singlet pair configurations, that is, any electron pair with total spin zero.

Specifically, the determinantal part is an antisymmetrized product of geminals (AGP), also called pairing functions:

$$\Phi_{\text{AS}} = \Phi_{\text{AGP}}(\mathbf{x}_1, \cdots, \mathbf{x}_n) = \hat{A}\left[\varphi(\mathbf{x}_1, \mathbf{x}_2), \cdots, \varphi(\mathbf{x}_{n-1}, \mathbf{x}_n)\right] \tag{2.65}$$

where $\hat{A}$ is an operator that symmetrize the product of the pairing functions. In our choice Φ_{AGP} can be written in a compact form as a determinant [119]:

$$\Phi_{\text{AGP}}(\mathbf{x}_1, \cdots, \mathbf{x}_n) = \det\left(A_{ij}\right) \tag{2.66}$$

where $\mathbf{A}$ is a $\frac{N}{2} \times \frac{N}{2}$ matrix of all the possible pairings:

$$A_{ij} = \varphi(\mathbf{x}_1, \mathbf{x}_2). \tag{2.67}$$

The geminals themselves are antisymmetric functions of two electrons coordinates written as the product of a spatial symmetric part and a spin singlet:

$$\varphi(\mathbf{x}_1, \mathbf{x}_2) = g(\mathbf{r}_1, \mathbf{r}_2)\frac{\delta(\sigma_i, \uparrow)\delta(\sigma_j, \downarrow) - \delta(\sigma_i, \downarrow)\delta(\sigma_j, \uparrow)}{\sqrt{2}}. \tag{2.68}$$

The spatial function $g(\mathbf{r}_1, \mathbf{r}_2)$ is defined starting from atom-centered basis as those in Eq.(2.8) and (2.9),

$$g(\mathbf{r}_i, \mathbf{r}_j) = \sum_{a,b}^{M} \sum_{v,u}^{N_b} \lambda_{vu}^{ab} \psi_{a,v}^{\mathrm{GTO}}(\mathbf{r}_i) \psi_{b,u}^{\mathrm{GTO}}(\mathbf{r}_j) \tag{2.69}$$

where the indices u and v collect the nlm indices of the GTOs for short. The matrix of parameters, $\mathbf{\Lambda} = \left\{\lambda_{u,v}^{a,b}\right\}$ gives, for fixed GTO channels u and v, the strength of the valence bond between atoms a and b, while for other atoms the parameters will vanish during the optimization phase.

If we diagonalizes the AGP matrix $\mathbf{\Lambda}$, the expression in Eq. 2.69 simplifies to

$$g(\mathbf{r}_i, \mathbf{r}_j) = \sum_{k}^{N_{\mathrm{MO}}} \lambda_k^{\mathrm{MO}} \chi_k^{\mathrm{MO}}(\mathbf{r}_i) \chi_k^{\mathrm{MO}}(\mathbf{r}_j) \tag{2.70}$$

where the product is now only between molecular orbitals. If only the first $N/2$ of them are retained, then the AGP matrix reduces to a Slater determinant wavefunction. One of the most important advantages of the AGP Ansatz is that it is equivalent to a linear combination of Slater determinants (i.e., multi-configurations), but the computational cost remains at the level of a single-determinant one. The multi-configurational nature of the AGP ansatz is what makes it suitable to take into account the static correlation.

Jastrow factor

The Jastrow factor is a function of the electron-electron and electron-ion distance, and as such it has multiple roles. First of all, it deals with the dynamic correlation of the electrons and it is fundamental to correctly describe the Van der Waals effects on the total energy [124], which are related to charge fluctuations. Secondly, it limits the double occupation of orbitals, accordingly with Pauli's exclusion principle. Finally, it ensures that the Kato's cusp conditions [125] is properly taken into account. The latter imposes that the wavefunction slope at nuclei position must have a cusp, a sharp change. Last, but not least, the presence of a Jastrow factor greatly accelerate the convergence in the parameters also in the antisymmetric part.

Considering the its exponential shape given in Eq. 2.64, the Jastrow exponent is the sum of three contributions:

$$J = J_1 + J_2 + J_3. \tag{2.71}$$

The one-body term itself is

$$J_1^{\mathrm{h}}(\mathbf{r}_1, \cdots, \mathbf{r}_N) = -\sum_{i}^{N} \sum_{a}^{M} (2Z_a)^{3/4} u\left((2Z_A)^{1/4} r_{ia}\right), \tag{2.72}$$

which satisfies the aforementioned Kato's cusp condition at electron-ion coalescence points, and where

$$u(|\mathbf{r}_i - \mathbf{q}_a|) = \frac{1 - e^{-b|\mathbf{r}_i - \mathbf{q}_a|}}{2b}; \tag{2.73}$$

is a simple bounded function. In both $\mathbf{r}_i$ and σ_i are the electron positions and spins respectively, $\mathbf{q}_a$ and Z_a are the atomic positions and number, l are the atomic orbitals indices assigned to a specific atom a.

In our specfic case, J_1 is applied only to hydrogen atoms, that are subjected to the bare Coulomb potential; in the case of oxygens the latter is replaced by the Burkatzki-Filippi-Dolg (BFD) potential [126].

In a similar way the two-body term manages the electron-electron cuspo conditions for antiparallel spin electrons

$$J_2 = \sum_{i<j}^{N} u(r_{ij}) \tag{2.74}$$

Finally the last term,

$$J_3 = \sum_{i<j}^{N} g(\mathbf{r}_i, \mathbf{r}_j), \tag{2.75}$$

includes many-body correlations through the use of geminals $g(\mathbf{r}_i, \mathbf{r}_j)$ as defined in 2.69, as they depend on the positions of two electrons i and j possibly belonging to two different atoms a and b.

2.3.4 Preparation and optimization of the quantum Monte Carlo wavefunction

In this Section we show the wavefunction specifications as reported in the SI of Ref. [127]

Preparation: geminal embedded orbitals

Before running finite-temperature calculations, we optimize a QMC variational wave function $\left|\Phi_{\mathbf{q}}\right\rangle$ at zero temperature.

Both Jastrow and AGP expansions are developed over a primitive O(3s2p1d) H(2s1p) and O(5s5p2d) H(4s2p) Gaussian basis functions, respectively. The primitive basis sets are then contracted using the geminal embedded orbitals (GEOs) scheme [128], reducing significantly the total number p of parameters describing the VMC wavefunction. This strategy is quite important to alleviate the computational burden of QMC, as in current optimization methods [117, 118, 129], which are based on iterative procedures that involve $p \times p$ matrices, the number of QMC samplings has to be much larger than p.

Previous works on the Zundel ion [130, 131] found that the optimal balance between accuracy and computational cost for the determinantal part is reached by the O[8]H[2] contracted GEO basis, in self-explaining notations. As the protonated water hexamer is a very similar system, in this work we used the same O[8]H[2] GEO contraction for the AGP part. Moreover, we further simplified the variational wavefunction previously developed for the Zundel ion, by contracting also the Jastrow basis set, using the same GEO embedding scheme. We tried different contraction sets, and tested them on the water dimer dissociation energy curve, as reported in Fig 2.1. The water dimer is a stringent benchmark for the quality of our wave function, as it has

a chemical complexity similar to the Zundel ion, with the main difference of being charge neutral. Charge neutrality allows us to directly probe the Jastrow capability of controlling charge fluctuations in the system, a fundamental property when coupled with the AGP determinantal part[132].

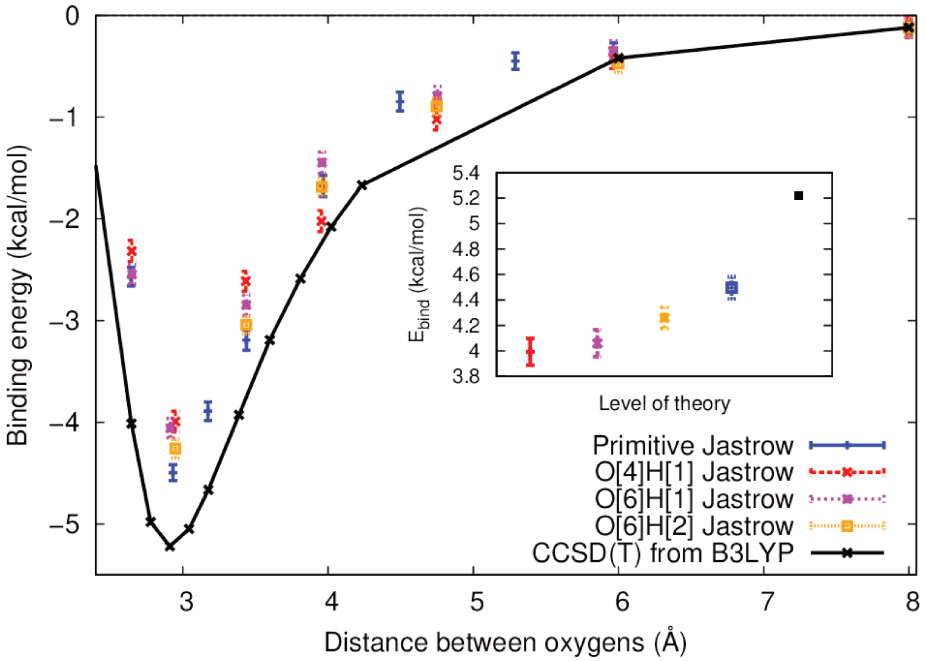


Figure 2.1: **Water dimer dissociation energy curve as a function of $d_{O_1O_2}$ obtained by VMC**. A Jastrow-Slater wave function has been employed, using different contracted basis sets in the Jastrow factor. Each trial wave function is built using the same basis set for the determinantal part, which is optimised together with the various Jastrow factors tested here. The black curve indicates the reference CCSD(T) result. From [127, 133].

As shown in Fig. 2.1, we find a systematic improvement as the number of GEOs orbitals increases, with the O[6]H[2] set yielding energies very close to the Jastrow primitive basis set reference at all oxygen-oxygen distances. As reported in Tab. 2.1, this is obtained with a number p of variational parameters significantly smaller than the one of the primitive basis set expansion. Thus, we used the O[6]H[2] GEO basis set for the Jastrow factor, and the O[8]H[2] GEO basis for the AGP part in all our subsequent molecular dynamics (MD) simulations of the protonated water hexamer. This results into a total number of 6418 variational parameters, comprising $g^{a,b}_{\mu,\nu}$, $\lambda^{a,b}_{\mu,\nu}$, the parameters of the homogeneous one-body and two-body Jastrow factors, and the linear coefficients of the Jastrow and determinantal basis sets (see Methods Section for a detailed description of the wave function parameters).

A more extended description of the variational wave function can be found in Ref. [130].

Optimization on the run

Once the optimal contraction scheme has been established, it is time to run the dynamics. To keep the simulation stable, the GTO exponents $\zeta_{l,n}$ (Eq. 2.9) in both the Jastrow and the AGP

Basis set	p	E_{bind} (kcal/mol)
Primitive Jastrow and primitive determinant	6303	4.46(8)
Primitive Jastrow and O[8]H[2] GEO determinant	2089	4.40(8)
O[6]H[2] GEO Jastrow and O[8]H[2] GEO determinant	1283	4.26(8)

Table 2.1: **Water dimer binding energies for QMC variational wave functions obtained with different types of basis set contractions.** The corresponding number p of variational parameters is also reported.

parts of the wavefunction are kept frozen. At each new ionic configuration, the wavefunction must be reoptimized with methods as the one introduced in Section 2.3.2. Since the ionic positions are smoothly connected to those of the previous MD time step, the electronic parameters will also evolve continuously. Therefore, only a few optimization steps are needed, especially in comparison with an wavefunction optimization from scratch.

Chapter 3

Ion dynamics

In this Chapter, we present the algorithms used for propagating the motion of the nuclei. Since they are built upon well-established frameworks, we also provide the broader context in which they are cast.

In Section 3.1, we introduce the formalism needed to sample observables at zero temperature, which is then adapted to the finite temperature case in Section 3.2 using stochastic differential equations. This framework is further extended to quantum simulations via the path integral formalism, as explained in Section 3.3. The specific algorithms employed in this work for classical simulations are the Bussi algorithm in presence of deterministic forces (Sec. 3.2.3), and the Attaccalite-Sorella algorithm in presence of QMC forces (Sec. 3.2.4). In the case of quantum simulations, both with deterministic and stochastically estimated forces, we used the Path integral Ornstein-Uhlenbeck dynamics for quantum simulations, described 3.5.2.

In this Chapter, since we are focusing solely on nuclei, unlike the previous chapter, we will denote the total number of atoms by N instead of M, as this notation is more customary in statistical mechanics.

3.1 Classical dynamics at zero temperature

3.1.1 Microcanonical ensemble and ergodicity

Consider a system of N *classical nuclei*, described by a set of degrees of freedom $\mathbf{\Gamma} = \{\mathbf{p}, \mathbf{q}\} = \{\mathbf{p}_a, \mathbf{q}_a\}_{a=1,\cdots,N}$. In the following we will often adopt the collective notation $\mathbf{p}$ and $\mathbf{q}$ for all the nuclei degrees of freedom, even when dealing with atoms with different masses. If the system is in thermodynamic equilibrium, it is known from statistical mechanics that a property $\mathcal{A}$, which is a function $A = A(\mathbf{\Gamma})$ of the degrees of freedom, can be derived by averaging it over the phase space according to the probability density function $\rho(\mathbf{\Gamma})$ associated with the ensemble taken into consideration:

$$\langle A(\mathbf{\Gamma}) \rangle = \int d\mathbf{\Gamma} \rho(\mathbf{\Gamma}) A(\mathbf{\Gamma}). \tag{3.1}$$

For a system with a constant number of particles N, volume V and energy E (NVE-ensemble,

or microcanonical ensemble) the probability density function is

$$\rho(\mathbf{\Gamma}) = \frac{1}{\Omega(N,V,E)}\delta\left(E - H(\mathbf{\Gamma})\right) \qquad E \leq H(\mathbf{\Gamma}) \leq E + \Delta \tag{3.2}$$

where Ω is the phase space volume corresponding to the shell of energy E:

$$\Omega = \int \delta\left(E - H(\mathbf{\Gamma})\right) \mathrm{d}\mathbf{\Gamma} = \int_{E \leq H(\mathbf{\Gamma}) \leq E+\Delta} \mathrm{d}\mathbf{\Gamma}. \tag{3.3}$$

Unfortunately only in a few cases the partition function can be evaluated analytically and it is necessary to resort to approximations, numerical methods or numerical simulations. The most common simulation methods are of two types: (i) Monte Carlo ones, in which the phase space is sampled according to the appropriate probability distribution function associated to the ensemble taken into consideration (direct computation of phase space or ensemble average); (ii) Molecular Dynamics (MD), in which the phase space is explored exploiting the dynamical equations of the system. For the latter method to be reliable, the dynamics must be *ergodic*, which means that time averages are equal to ensemble averages in the limit $T \to \infty$:

$$\langle A(\mathbf{\Gamma}) \rangle = \lim_{T\to\infty} \frac{1}{T} \int_0^T dt A(\mathbf{\Gamma}(t)). \tag{3.4}$$

This computation is done in practice with a finite time step Δt and in a definite interval of time $T = N_{\text{step}}\Delta t$ of simulation:

$$\langle A(\mathbf{\Gamma}) \rangle \approx \frac{1}{N_{\text{step}}\Delta t} \sum_{i=1}^{N_{\text{step}}} A(\mathbf{\Gamma}(t_i)), \tag{3.5}$$

where the length of the simulation must be long enough in order to satisfy two conditions: the phase space is explored as much as possible; all the phenomena at different time scales are reproduced correctly.

3.1.2 Time evolution via Liouvillian operator

How is the system propagated in time? In Chapter 1 we started from a full quantum problem, separating nuclei and electrons wavefunctions by means of the BO approximation. By writing the nuclear wave function in a quantum fluid dynamics representation it is possible to derive the classical equation of motion of the nuclei [39], which in its Newtonian form reads:

$$m_a \ddot{\mathbf{q}}_a = -\nabla_a E_0(\mathbf{q}) = \mathbf{f}_a. \tag{3.6}$$

where E_0 is the energy eigenvalue of the electronic Hamiltonian. This means that once we know the solution to the eigenvalue problem for the electrons, i.e. once we know PES $E_0(\mathbf{q})$, already introduced in Chapter 1, with any of the electronic structure methods described in Chapter 2, by deriving it with respect to the coordinates of a single nucleus, $\mathbf{q}_a$, we obtain the force needed to propagate it [134]. For this reason we can say that classical nuclei moves *on* the PES generated by the electrons, which is explored and characterized during the MD simulation.

In the context of the Hamiltonian formulation of classical mechanics, Equation (3.6) can be studied in the Liouville formalism, which is particularly useful for formulating propagation algorithms. The time evolution of momenta and positions can be rewritten as:

$$\begin{cases} \dot{\mathbf{p}}_a = -\nabla_{\mathbf{q}_a} H = \mathbf{f}_a \\ \dot{\mathbf{q}}_a = \nabla_{\mathbf{p}_a} H = \frac{\mathbf{p}_a}{m_a} \end{cases} \quad \leftrightarrow \frac{\partial \mathbf{\Gamma}}{\partial t} = -\{H, \mathbf{\Gamma}\}, \tag{3.7}$$

where we used the Poisson bracket, which allows one to conveniently express the Liouville operator in a compact form. Within this formalism, the Liouville operatore can be defined as

$$i\hat{L} \equiv \nabla_{\mathbf{p}} H \nabla_{\mathbf{q}} - \nabla_{\mathbf{q}} H \nabla_{\mathbf{p}} = -\{H, \cdot\}, \tag{3.8}$$

so that the equation of motion and its formal solution read:

$$\frac{\mathrm{d}\mathbf{\Gamma}}{\mathrm{d}t} = i\hat{L}\mathbf{\Gamma}, \tag{3.9}$$

$$\mathbf{\Gamma}(t) = e^{i\hat{L}t}\mathbf{\Gamma}(0) = e^{i(\hat{L}_{\mathbf{p}} + \hat{L}_{\mathbf{q}})t}\mathbf{\Gamma}(0), \tag{3.10}$$

respectively. In the last Equation, the Liouvillian is expressed as the sum of two terms, namely $i\hat{L} = i\hat{L}_{\mathbf{p}} + i\hat{L}_{\mathbf{q}}$, where:

$$i\hat{L}_{\mathbf{q}} = \dot{\mathbf{q}} \cdot \nabla_{\mathbf{q}} = \frac{\mathbf{p}}{m} \cdot \nabla_{\mathbf{q}}, \tag{3.11}$$

$$i\hat{L}_{\mathbf{p}} = \dot{\mathbf{p}} \cdot \nabla_{\mathbf{p}} = \mathbf{f} \cdot \nabla_{\mathbf{p}}. \tag{3.12}$$

This formalism will reveal useful in establishing a common framework for the definition of various molecular dynamics algorithms.

3.1.3 Velocity-Verlet algorithm

The Verlet algorithm [135] is one of the simplest and most employed integration schemes since its conception at the end of the 1960s. Being the starting point of many more sophisticated methods as the ones showed later, we briefly remind its steps in its velocity variant [136].

The exponential that appears in Eq. 3.10 is approximatated according to the Suzuki-Trotter second order decomposition [137, 138]:

$$e^{i\hat{L}_{\mathbf{p}}\delta t/2} e^{i\hat{L}_{\mathbf{q}}\delta t} e^{i\hat{L}_{\mathbf{p}}\delta t/2} + \mathcal{O}(\delta t^2). \tag{3.13}$$

Notice that if we were dealing with numbers at the exponent, the above Equation would be exact without the need of the $\mathcal{O}(\delta t^2)$ term. However, here we are dealing with operators which in general do not commute with each other. Therefore, the product of the exponentials is an approximations up to order $O(\delta t^2)$.
After the Suzuki-Trotter break-up, the velocity-Verlet steps are the following:

1. Propagate the particle momenta for $\delta t/2$

$$\mathbf{p}(t + \delta t/2) = e^{i\hat{L}_{\mathbf{p}}\delta t/2}\mathbf{p}(t) = \left(1 + \frac{\delta t}{2}\mathbf{f} \cdot \nabla_{\mathbf{p}}\right)\mathbf{p}(t) = \mathbf{p}(t) + \frac{\delta t}{2}\mathbf{f}(t). \tag{3.14}$$

2. Propagate the ionic positions for δt,

$$\mathbf{q}(t+\delta t) = e^{i\hat{L}_{\mathbf{q}}\delta t}\mathbf{q}(t) = \left(1 + \delta t\frac{\mathbf{p}(t)}{m}\cdot\nabla_{\mathbf{q}}\right)\mathbf{q}(t) = \mathbf{q}(t)+\delta t\frac{\mathbf{p}(t+\delta t/2)}{m} = \mathbf{q}(t)+\delta t\frac{\mathbf{p}(t)}{m}+\delta t^2\frac{\mathbf{f}(t)}{2m}. \tag{3.15}$$

3. Evaluate the Born-Oppenheimer forces in the new positions (in our case, using *ab initio* methods or machine learning potentials):

$$\mathbf{f}(t+\delta t) = -\nabla_{\mathbf{q}}E_0\left(\mathbf{q}(t+\delta t)\right). \tag{3.16}$$

4. Propagate the particle momenta for the remaining half time step, from $t+\delta t/2$ to $t+\delta t$:

$$\mathbf{p}(t+\delta t) = e^{i\hat{L}_{\mathbf{p}}\delta t/2}\mathbf{p}(t+\delta t/2) = \mathbf{p}(t+\delta t/2) + \frac{\delta t}{2}\mathbf{f}(t+\delta t). \tag{3.17}$$

Notice that the only approximation is the Suzuki-Trotter breakup: the development of the exponential up to the first order, as in Eq. 3.14, 3.15 and 3.17, based on the small time step δt, is exact, since terms involving powers of the gradients $\nabla_{\mathbf{p}}$ and $\nabla_{\mathbf{q}}$ are zero.

3.2 Classical dynamics at finite temperature

3.2.1 Canonical ensemble

If we are interested in properties which depend on the temperature, we must look at a different ensemble, namely the *canonical* one. In this case the probability density is different from the one (3.2):

$$\rho(\mathbf{\Gamma}) = \frac{e^{-\beta H(\mathbf{\Gamma})}}{Z(N,V,T)}, \tag{3.18}$$

where we have the usual prefactor $\beta = 1/k_BT$ and the partition function is defined as:

$$Z(N,V,T) = \int \mathrm{d}\mathbf{\Gamma} e^{-\beta H(\mathbf{\Gamma})}. \tag{3.19}$$

As a final remark we remind that Z is related to the Helmholtz free energy $F(N,V,T) = U - TS$ via:

$$F(N,V,T) = -k_BT\ln Z(N,V,T). \tag{3.20}$$

3.2.2 Thermostatting by Langevin dynamics

In order to sample the right distribution function (3.18) a number of thermostatting schemes has been developed. They can be divided in two categories:

1. *Deterministic thermostats*, which correct the velocities of the particles in order to keep the system at constant temperature. Among them, the Nosé-Hoover [139, 140] is one of the most known.

2. *Stochastic thermostats*, which treat the particles as Brownian ones, subjected to a dissipative force and a stochastic force, in addition to the external one, such that the constant temperature condition is satisfied via the Fluctuation-Dissipation Theorem (FDT).

The second type of thermostats is suitable in dealing with the stochastic nature of forces generated by quantum Monte Carlo methods, therefore we will proceed in explaining their fundamentals features.

Stochastic thermostats in MD simulation are usually built upon Langevin equations, a class of Stochastic Differential Equations (SDEs) originally conceived to describe the random motion of a mesoscopic particle immersed in a thermal bath, that is a Brownian particle. Despite its historical origin, the Langevin approach paved the way to an entire new field of stochastic processes [141] and their applications to different natural phenomena and algorithms, including the molecular dynamics thermostatting. In fact we are not dealing with a particle in a thermal bath, but with an isolated system. Nevertheless we can employ the Langevin equation to impose a dynamics at a fixed temperature by adding an opportune white noise.

In its under-damped differential form, the Langevin equation reads[1]

$$\begin{aligned} \dot{\mathbf{p}}(t) &= -\overline{\gamma}\mathbf{p}(t) + \mathbf{f}\left(\mathbf{q}(t)\right) + \boldsymbol{\eta}(t) \\ \dot{\mathbf{q}}(t) &= \frac{\mathbf{p}}{m}, \end{aligned} \tag{3.24}$$

where at each step the random force $\boldsymbol{\eta}(t)$ is a random vector sampled from a multivariate Gaussian white noise distribution $\mathcal{N}(0, \mathrm{d}t)$. As such, these random vectors must satisfy the zero mean condition

$$\left\langle \boldsymbol{\eta}(t) \right\rangle = \mathbf{0}, \tag{3.25}$$

and they must be statistically independent in time

$$\left\langle \boldsymbol{\eta}(t)\boldsymbol{\eta}^T(t') \right\rangle = \delta(t - t')\overline{\boldsymbol{\alpha}}(\mathbf{q}). \tag{3.26}$$

The latter condition translates into the fluctuation-dissipation theorem, which relates the covariance matrix of the stochastic forces, $\overline{\boldsymbol{\alpha}}(\mathbf{q})$, to the friction matrix $\overline{\gamma}(\mathbf{q})$, also called *damping matrix*,

[1]The rigorous way of writing Eq. (3.24) is by using stochastic differentials:

$$\begin{aligned} \mathrm{d}\mathbf{p}(t) &= -\overline{\gamma}(\mathbf{q})\mathbf{p}(t)\,\mathrm{d}t + \mathbf{f}\left(\mathbf{q}(t)\right)\mathrm{d}t + \mathbf{B}\left(\mathbf{q}(t)\right)\mathrm{d}\mathbf{W}(t) \\ \mathrm{d}\mathbf{q}(t) &= \frac{\mathbf{p}}{m}\,\mathrm{d}t, \end{aligned} \tag{3.21}$$

where $\mathrm{d}\mathbf{W}(t)$ is a Wiener process, a continuous but non-differentiable function of time. Nevertheless, we can formally define the Gaussian white noise as:

$$''\boldsymbol{\eta}(t) = \lim_{\mathrm{d}t\to 0} \mathbf{B}\frac{\mathrm{d}\mathbf{W}(t)}{\mathrm{d}t}''. \tag{3.22}$$

Heuristically, we can say that $\mathrm{d}\mathbf{W} \approx (\mathrm{d}t)^{1/2}$, which distinguishes the time dependence of Brownian motion from that of typical ballistic motion.

The solution of Eq. (3.21) is

$$\mathbf{p}(t) = p_0 - \int_0^t \mathrm{d}s\left(\overline{\gamma}(\mathbf{q})\mathbf{p}(t) + \mathbf{f}(\mathbf{q})\right) + \int_0^t \mathrm{d}s\mathbf{B}(\mathbf{q})\left(\mathbf{q}(t)\right)\mathrm{d}\mathbf{W}(t). \tag{3.23}$$

Despite its familiarity, solving such integrals requires a completely different and fascinating way of doing calculus, for which the reader is referred to [141, 142]. In this exposition, we stick to the physicists dot notation.

through a relation that involves the temperature T [143]:

$$2mk_BT\overline{\gamma}(\mathbf{q}) = \overline{\boldsymbol{\alpha}}(\mathbf{q}). \tag{3.27}$$

Notice that Eq. 3.24 is just the Newtonian formula to which we added the dissipation and the fluctuation terms. As we did in Section 3.1.2, we can introduce a formalism to deal with the probability distribution of an ensemble defined on the phase space. Indeed, it is well known from theory that SDE problems can be treated from a macroscopic perspective using generalized Fokker-Planck equation (FPE) [142].

If the particle were subjected to drift and diffusion forces only, with the drift depending linearly on the momenta, we would have an Ornstein-Uhlenbeck (OU) process [144] in the momenta space:

$$\dot{\mathbf{p}}(t) = -\gamma\mathbf{p}(t) + \boldsymbol{\eta}(t) \tag{3.28}$$

and the probability distribution $\rho(\mathbf{p}, \mathbf{q}, t)$ would evolve in time according to the following FPE:

$$\frac{\partial\rho(\mathbf{p},\mathbf{q},t)}{\partial t} = -\overline{\gamma}\underbrace{\left(\nabla_{\mathbf{p}}\mathbf{p}\right)}_{\text{drift}} + \underbrace{\frac{m}{\beta}\nabla^2_{\mathbf{p}}}_{\text{diffusion}}\rho(\mathbf{p},\mathbf{q},t) = -i\hat{L}_{\text{FP}}\rho(\mathbf{p},\mathbf{q},t), \tag{3.29}$$

where we have defined the *Fokker-Planck operator* $\hat{L}_{\text{FP}}$.

In the more general case including also the action of an external force $\mathbf{f}$, like in Eq. (3.24), we consider the *Kramers-Klein operator* instead:

$$i\hat{L}_{\text{KK}} = i\hat{L}_{\mathbf{p}} + i\hat{L}_{\mathbf{q}} + i\hat{L}_{\text{FP}} \tag{3.30}$$

which resembles the decomposition of the Liouvillian operator reported in Eq. 3.10, 3.11 and 3.12.

3.2.3 Bussi algorithm

The Bussi algorithm [145] can be described by the Suzuki-Trotter decomposition of operators in Eq. 3.30:

$$e^{i\hat{L}_{\text{KK}}\delta t} = e^{\left(i\hat{L}_{\mathbf{p}}+i\hat{L}_{\mathbf{q}}+i\hat{L}_{\text{FP}}\right)\delta t} \approx e^{i\hat{L}_{\text{FP}}\delta t/2}\underbrace{e^{i\hat{L}_{\mathbf{p}}\delta t/2}e^{i\hat{L}_{\mathbf{q}}\delta t}e^{i\hat{L}_{\mathbf{p}}\delta t/2}}_{\text{velocity-Verlet}}e^{i\hat{L}_{\text{FP}}\delta t/2}, \tag{3.31}$$

where we can see we have a deterministic propagation step analogous to the velocity-Verlet algorithm, sandwiched between two stochastic propagation steps based on the OU-process Equation (3.28), where the BO forces do not act. It is possible to compute the exact thermostat propagation for any time interval Δt [146]. This derivation can be found in Appendix A.

The whole propagation according to Eq. (3.31) comprises the following steps:

1. First *analytical* thermostatting of particle momenta

$$\mathbf{p}(t^+) = c_1\mathbf{p}(t) + c_2\mathbf{R}(t). \tag{3.32}$$

2. Deterministic approximate propagation according to the Verlet algorithm:

- Propagate the positions for an entire time-step δt

$$\mathbf{q}(t+\delta t) = e^{i\hat{L}_{\mathbf{q}}\delta t} = \mathbf{q}(t) + \delta t \frac{\mathbf{p}(t^+)}{m} + \delta t^2 \frac{\mathbf{f}(t)}{2m}. \tag{3.33}$$

- Compute the new BO forces according to the new configuration

$$\mathbf{f}(t+\delta t) = -\nabla_{\mathbf{q}} V\left(\mathbf{q}(t+\delta t)\right). \tag{3.34}$$

- Approximate the momenta given the old and the new forces

$$\mathbf{p}(t^- + \delta t) = e^{i\hat{L}_{\mathbf{p}}\delta t}\mathbf{p}(t^+) = \mathbf{p}(t^+) + \frac{\mathbf{f}(t) + \mathbf{f}(t+\delta t)}{2}\delta t. \tag{3.35}$$

3. Last *analytical* thermostatting of particle momenta

$$\mathbf{p}(t+\delta t) = c_1 \mathbf{p}(t^- + \delta t) + c_2 \mathbf{R}(t+\delta t). \tag{3.36}$$

In all the thermostatting steps the coefficients, $\mathbf{R}$ is a Gaussian random vector, while c_1 and c_2 are:

$$c_1 = e^{-\overline{\gamma}\frac{\delta t}{2}} \qquad c_2 = \sqrt{\left(1 - e^{-\overline{\gamma}\delta t}\right)\frac{m}{\beta}}, \tag{3.37}$$

as motivated in Appendix A. The timestamps t^+ and t^- refer to the instants of time just after and just before the application of the thermostat to the momenta, respectively. More precisely, $p(t^+)$ in Eq. (3.32) are the momenta thermostatted for $\delta t/2$; still, these are not the fully propagated momenta yet, because the BO forces will act in the following step, according to the algorithm. Analogously, $p(t^- + \delta t)$ in Eq. (3.36) are the momenta that still miss the last half-contribution of the thermostat.

The limitation of this algorithm is that it can not deal with Born-Oppenheimer forces $\mathbf{f}$ intrinsically affected by a stochastic noise, such as those computed through QMC. The latter would add more noise to the integration scheme, increasing the effective temperature of the simulation.

3.2.4 Attaccalite-Sorella algorithm

The solutions to dynamics biased toward higher temperatures are based on noise covariance correction schemes. Before introducing these, it is useful to switch to transformed variables:

$$\begin{aligned} \mathbf{q} &= \mathbf{q}_0\sqrt{m} \\ \mathbf{p} &= \mathbf{p}_0/\sqrt{m} \\ \boldsymbol{\eta} &= \boldsymbol{\eta}_0\sqrt{m} \\ \mathbf{f} &= \mathbf{f}_0/\sqrt{m}, \end{aligned} \tag{3.38}$$

where the variables indexed by zeroes are the original coordinates. By applying this transformation to the variables involved in the Langevin equation 3.24, we would get:

$$\begin{aligned} \dot{\mathbf{p}}(t) &= -\overline{\gamma}(\mathbf{q})\mathbf{p}(t) + \mathbf{f}\left(\mathbf{q}(t)\right) + \boldsymbol{\eta}(t) \\ \dot{\mathbf{q}}(t) &= \mathbf{p}, \end{aligned} \tag{3.39}$$

with the noise mean and covariance relations rescaled as:

$$\langle \boldsymbol{\eta}(t) \rangle = \mathbf{0}, \qquad \left\langle \boldsymbol{\eta}(t)\boldsymbol{\eta}^T(t') \right\rangle = \overline{\boldsymbol{\alpha}}(\mathbf{q})\delta(t-t') = 2k_B T\overline{\boldsymbol{\gamma}}(\mathbf{q})\delta(t-t'), \tag{3.40}$$

where in the last step we used fluctuation-dissipation relation without masses, at variance with Eq. (3.27). Notice that the covariance matrix $\overline{\boldsymbol{\alpha}}$ deals with two types of correlation:

- *Spatial correlation*, or cross-correlation, between the vectorial forces components.
- *Time correlation*, which in the Markovian case is reduced to a δ-function, in order to have Gaussian white noise; otherwise we would have *colored noise*.

Time discretization approximation

The noise correction schemes, introduced by Attaccalite-Sorella (AS) in Ref. [112], and later refined in Refs. [147, 148], are all based on the time discretization approximation.

We introduce it by first showing the formal solution of the rescaled Langevin Equation 3.39, found by integrating from time t to time t':

$$\begin{aligned} \mathbf{p}(t') - \mathbf{p}(t) &= \int_t^{t'} \mathrm{d}s \left(-\overline{\gamma}_{\mathbf{q}}\mathbf{p}(s) + \mathbf{f}_{\mathbf{q}(s)} + \boldsymbol{\eta}(s) \right), \\ \mathbf{q}(t') - \mathbf{q}(t) &= (t'-t)\mathbf{p}. \end{aligned} \tag{3.41}$$

Without further information, or approximations, the integration of the first equation in momenta variables can be developped up to the following form:

$$\mathbf{p}(t') = \mathbf{p}(t)e^{-\int_t^{t'} \mathrm{d}s\overline{\gamma}_{\mathbf{q}(s)}} + \int_t^{t'} \mathrm{d}s e^{-\overline{\gamma}_{\mathbf{q}(s)}(t'-s)} \left(\mathbf{f}_{\mathbf{q}(s)} + \boldsymbol{\eta}(s) \right), \tag{3.42}$$

where for readability we expressed the $\mathbf{q}(s)$-dependences in subscript.

To further develop the solution, the time is discretized in small time intervals[2], $t' - t = \delta t$, and in each of these interval the dependence of forces and friction matrix on the positions $\mathbf{q}$ is neglected, resulting in the following constant values

$$\begin{aligned} \mathbf{f}(\mathbf{q}(s)) &= \mathbf{f}\left(\mathbf{q}(t_n)\right) \approx \mathbf{f}_n \\ \overline{\boldsymbol{\gamma}}(\mathbf{q}(s)) &= \overline{\boldsymbol{\gamma}}\left(\mathbf{q}(t_n)\right) \approx \overline{\boldsymbol{\gamma}}_n, \end{aligned} \tag{3.43}$$

that can be easily put outside of the integrals appearing in Eq. 3.42

$$(3.42) \approx \mathbf{p}(t)e^{-\overline{\gamma}_n\delta t} + \overline{\gamma}_n^{-1}\left(1 - e^{-\overline{\gamma}_n\delta t}\right)(\mathbf{f}_n + \boldsymbol{\eta}_n) = \mathbf{p}(t)e^{-\overline{\gamma}_n\delta t} + \overline{\boldsymbol{\Gamma}}_n(\mathbf{f}_n + \boldsymbol{\eta}_n). \tag{3.44}$$

where we defined $\overline{\boldsymbol{\Gamma}}_n = \overline{\gamma}_n^{-1}\left(1 - e^{-\overline{\gamma}_n\delta t}\right)$. Then the solution of the scaled Langevin Eq. 3.39 can be approximated as:

$$\begin{aligned} \mathbf{p}_{n+1} &= \mathbf{p}_n e^{-\overline{\gamma}_n\delta t} + \overline{\boldsymbol{\Gamma}}_n(\mathbf{f}_n + \boldsymbol{\eta}_n) \\ \mathbf{q}_{n+1} &= \mathbf{q}_n + \mathbf{p}_n\delta t \end{aligned} \tag{3.45}$$

[2]Notice that the introduction of this small timestep is a true approximation, that has nothing to do with the stochastic differential formulation of SDE (Eq. 3.21).

where we expressed the dependence on the discrete time t_n directly putting n as subscript for short. The term $\boldsymbol{\eta}_n$ is the noise vector obtained from integration of the GWN $\boldsymbol{\eta}(t)$ in the time interval $[t_n - \delta t/2, t_n + \delta t/2]$:

$$\boldsymbol{\eta}_n = \frac{\overline{\boldsymbol{\gamma}}_n}{2\sinh(\overline{\boldsymbol{\gamma}}_n \delta t/2)} \int_{t_n-\delta t/2}^{t_n+\delta t/2} \mathrm{d}s \boldsymbol{\eta}(s) e^{-\overline{\boldsymbol{\gamma}}_n (t_n - s)}, \tag{3.46}$$

where the prefactor outside of the integral accounts for the multiplication by $\overline{\boldsymbol{\Gamma}}_n$ in Eq. 3.45.

The integrated noise is characterized by the following covariance matrix:

$$\left\langle \boldsymbol{\eta}_m \boldsymbol{\eta}_n^T \right\rangle = \frac{\overline{\boldsymbol{\gamma}}_m \overline{\boldsymbol{\gamma}}_n}{4\sinh(\overline{\boldsymbol{\gamma}}_m \delta t/2)\sinh(\overline{\boldsymbol{\gamma}}_n \delta t/2)} \int_{t_m-\delta t/2}^{t_m+\delta t/2} \int_{t_n-\delta t/2}^{t_n+\delta t/2} e^{-\overline{\boldsymbol{\gamma}}_m (t_n - r)} e^{-\overline{\boldsymbol{\gamma}}_n (t_n - s)} \left\langle \boldsymbol{\eta}(r) \boldsymbol{\eta}^T(s) \right\rangle. \tag{3.47}$$

We know from Eq. 3.40 the covariance in the integral is nonzero only when the two noises are evaluated at the same time, $r = s$, which implies that also the time intervals must be the same, $t_m = t_n$. Therefore, the covariance matrix reduces to $\left\langle \boldsymbol{\eta}_n^T \boldsymbol{\eta}_n \right\rangle$

$$\left\langle \boldsymbol{\eta}_m \boldsymbol{\eta}_n^T \right\rangle \delta_{mn} = \left\langle \boldsymbol{\eta}_n \boldsymbol{\eta}_n^T \right\rangle = \frac{\overline{\boldsymbol{\gamma}}_n^2}{4\sinh^2(\delta t/2)} \left(\int_{t_n-\delta t/2}^{t_n+\delta t/2} e^{-\overline{\boldsymbol{\gamma}}_n (t_n - s)} \, \mathrm{d}t \right)^2 = k_B T \overline{\boldsymbol{\gamma}}_n^2 \coth\left(\overline{\boldsymbol{\gamma}}_n \frac{\delta t}{2} \right). \tag{3.48}$$

Noise correction

Knowing that the whole noise added by the thermostat should have a covariance matrix as the one in Eq. 3.48, we can deduce a noise correction scheme where the actual random forces, $\boldsymbol{\eta}_{\mathrm{ext},n}$, are sampled according to a multivariate Gaussian distribution having a noise-corrected covariance matrix:

$$\left\langle \boldsymbol{\eta}_{\mathrm{ext},n} \boldsymbol{\eta}_{\mathrm{ext},n}^T \right\rangle = \left\langle \boldsymbol{\eta}_n \boldsymbol{\eta}_n^T \right\rangle - \left\langle \delta \mathbf{f}_n \delta \mathbf{f}_n^T \right\rangle \tag{3.49}$$

where $\left\langle \boldsymbol{\eta}_n \boldsymbol{\eta}_n^T \right\rangle$ is determined as in Eq. (3.48), and the last term is the (integrated) QMC forces covariance matrix, $\overline{\boldsymbol{\alpha}}^{\mathrm{QMC}}(\mathbf{q}) = \langle \delta \mathbf{f}(\mathbf{q}) \delta \mathbf{f}^T(\mathbf{q}) \rangle$. Equation 3.49 represents the core of the AS algorithm.

Considering that the stochastic nature of QMC forces introduce additional spatially-correlated noise at each time-step of the dynamics, a further development consists in optimizing the value of the $\overline{\boldsymbol{\gamma}}$ matrix (which in this approach has non trivial off-diagonal matrix elements) by choosing it such that $\gamma = \bar{\alpha}/2k_B T$, where the stochastic forces-covariance matrix reads as:

$$\overline{\boldsymbol{\alpha}}(\mathbf{q}) = \alpha_0 \bar{\mathbf{I}} + \Delta_0 \overline{\boldsymbol{\alpha}}^{\mathrm{QMC}}(\mathbf{q}). \tag{3.50}$$

In the above Equation, the first term of the sum is $\alpha_0 = 2k_B T\gamma$, the diagonal white noise contribution whose parameter γ is selected by the user, and Δ_0 is an additional user-tunable parameter to make the covariance matrix positive definite and, together with γ, to lead to an optimal damping matrix $\overline{\boldsymbol{\gamma}}$ and thus a more efficient Langevin dynamics.

3.2.5 Classical Momentum-Position Correlator

The AS algorithm has later been improved with the Classical momentum-position correlator (CMPC) algorithm [148, 149], in which the noise due to Langevin dynamics affects both positions and momenta. To account for this, new $6N$-dimensional vectors that combines momenta and positions coordinates, ionic forces and random forces vector are defined as:

$$\mathbf{X} = \begin{pmatrix} \mathbf{p} \\ \mathbf{q} \end{pmatrix}, \qquad \mathbf{F} = \begin{pmatrix} \mathbf{f} \\ \mathbf{0} \end{pmatrix}, \qquad \boldsymbol{\Xi} = \begin{pmatrix} \boldsymbol{\eta} \\ \mathbf{0} \end{pmatrix}, \tag{3.51}$$

respectively. These variables allow to write the Langevin Eq. (3.39) as

$$\dot{\mathbf{X}} = -\hat{\overline{\gamma}}\mathbf{X} + \mathbf{F} + \boldsymbol{\Xi}, \tag{3.52}$$

where the $6N \times 6N$ matrix $\hat{\overline{\gamma}}$ represents a generalized friction that couples both momenta and positions:

$$\hat{\overline{\gamma}} = \begin{pmatrix} \overline{\gamma} & \mathbf{0} \\ -\mathbf{I} & \mathbf{0} \end{pmatrix}, \tag{3.53}$$

where $\overline{\gamma}$ being the same $3N \times 3N$ friction matrix defined in Eq. (3.40), and $\mathbf{I}$ is the identity matrix. The formal solution of Eq. (3.52) is similar to the one obtained in Eq. (3.42)

$$\mathbf{X}(t') = e^{-\hat{\overline{\gamma}}(t'-t)}\mathbf{X}(t) + \int_t^{t'} \mathrm{d}s e^{\hat{\overline{\gamma}}(s-t')} \left(\mathbf{F}(\mathbf{X}(s)) + \boldsymbol{\Xi}(s)\right). \tag{3.54}$$

If we express $e^{-\hat{\overline{\gamma}}\delta t}$ in terms of Pauli matrices

$$\hat{\overline{\gamma}} = \frac{\overline{\gamma}}{2} \otimes \mathbf{I} - \frac{\mathbf{I}}{2} \otimes \sigma_x + i\frac{\mathbf{I}}{2} \otimes \sigma_y + \frac{\overline{\gamma}}{2} \otimes \sigma_z, \tag{3.55}$$

it is possible to express the solution Eq. (3.54) in a closed form:

$$\mathbf{p}_{n+1} = e^{-\overline{\gamma}\delta t}\mathbf{p}_n + \boldsymbol{\Gamma}\left(\mathbf{f}_n + \tilde{\boldsymbol{\eta}}\right) \tag{3.56}$$

$$\mathbf{q}_{n+1} = \mathbf{q}_n + \overline{\gamma}\mathbf{p}_n + \boldsymbol{\Theta}\left(\mathbf{f}_n + \tilde{\tilde{\boldsymbol{\eta}}}\right), \tag{3.57}$$

where the time evolution has been discretized with timestep δt, and the subscripts refer to the corresponding time interval, such that $\mathbf{p}_n = \mathbf{p}(t_n)$, $\mathbf{q}_n = \mathbf{q}(t_n)$, and $\mathbf{f}_n = \mathbf{f}(\mathbf{q}(t_n))$. Notice that also here we assumed that $\mathbf{f}_n$ and $\overline{\gamma}_n$ do not vary withing the small time interval. The remaining symbols in the above equations are defined as

$$\begin{aligned}
\boldsymbol{\Gamma} &= \overline{\gamma}^{-1}(1 - e^{-\overline{\gamma}\delta t}), \\
\boldsymbol{\Theta} &= \overline{\gamma}^{-2}(-1 + \overline{\gamma}\delta t + e^{-\overline{\gamma}\delta t}), \\
\tilde{\boldsymbol{\eta}} &= \boldsymbol{\Gamma}^{-1} \int_{t_n}^{t_{n+1}} \mathrm{d}t e^{\overline{\gamma}(t-t_{n+1})}\boldsymbol{\eta}(t), \\
\tilde{\tilde{\boldsymbol{\eta}}} &= (\boldsymbol{\Theta}\overline{\gamma})^{-1} \int_{t_n}^{t_{n+1}} \mathrm{d}t (1 - e^{\overline{\gamma}(t-t_{n+1})})\boldsymbol{\eta}(t).
\end{aligned} \tag{3.58}$$

The strength of CMPC algorithm is that it propagates momenta and positions *simultaneously* in a single iteration thanks to the use of momentum-position correlation matrices. In particular, according to Eqs. (3.56)) and (3.57)), not only the momenta but also the positions are affected by the integrated Langevin noise.

While more cumbersome, this derivation is useful because it introduces some ideas used to derive the PIOUD algorithm in the PIMD formalism (see Section 3.5.2).

3.3 Quantum dynamics in the path integral formalism

3.3.1 Nuclear quantum effects

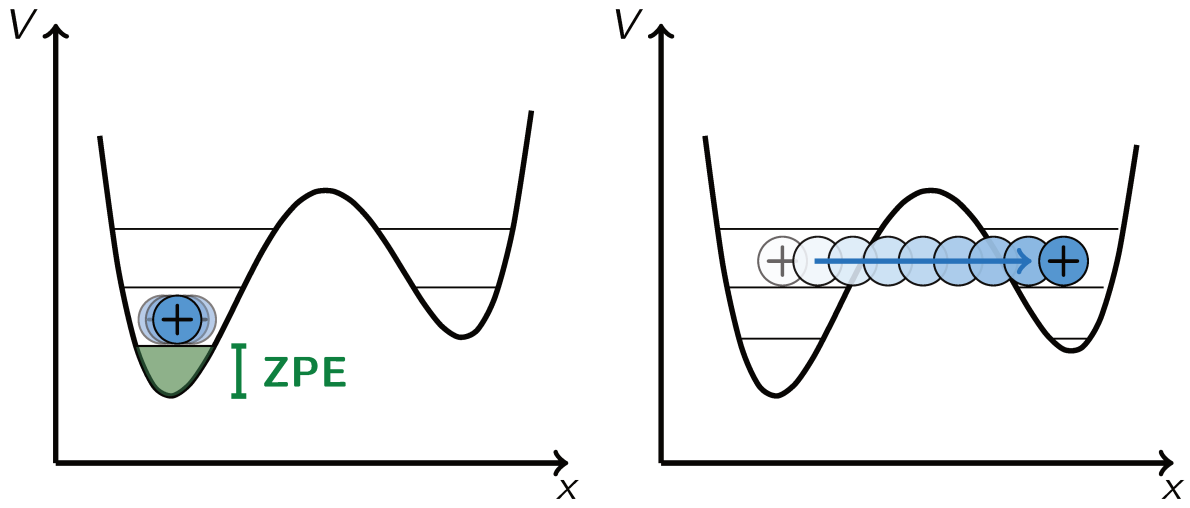
Nuclear Quantum Effects (NQEs) are the manifestation of the quantum nature of nuclei on chemical properties that cannot be fully explained by classical physics. These effects are particularly significant for light nuclei, which can exhibit behaviors such as Zero Point Energy (ZPE), proton delocalization, energy level quantization, and proton tunneling.

ZPE and proton delocalization arise directly from Heisenberg's uncertainty principle, which states that a quantum system will always have finite energy, even in its lowest energy state, and that its exact position cannot be precisely determined. In molecular systems, ZPE is defined as the energy difference between the vibrational ground state and the minimum of the Born-Oppenheimer potential energy surface. It can be estimated as $\hbar\omega_0/2$, where $\hbar = h/2\pi$ is the reduced Planck's constant and ω_0 is the frequency of the lowest vibrational mode.

Tunneling refers to the non-zero probability of a particle crossing an energy barrier without needing thermal fluctuations, a process that is classically forbidden. This phenomenon depends on both the energy scale involved and the mass of the particle; for example, tunneling is much more likely for electrons [150] than for protons [28].

Due to the presence of hydrogen, NQEs have been observed in water systems [32], and they play a crucial role in protonated water as well. In particular, NQEs affect Hydrogen bond (H-bond) and the proton transfer (PT) mechanism [151], influencing the reaction rates of PT [152].

In the case of PT along H-bond, an oversimplified but illustrative model of how quantum nuclei might behave can be described by reducing the full PES to a double-well potential (Fig. 3.1). Taking the example of water, the double potential well is formed between the two oxygen atoms sharing a proton, with the H^+ positioned in one of the wells. The central barrier and the symmetry of the well change as the oxygen-oxygen distance OO varies: when the oxygens are pulled farther apart, the barrier height increases, and the double well becomes asymmetric, as the proton tends to form a covalent bond with the nearest oxygen. Conversely, when OO decreases, the barrier is lowered, and the potential well becomes more symmetric, making proton transfer more feasible.



(a) Delocalization and zero point energy (b) Proton tunneling

Figure 3.1: **Cartoonish pictures of NQEs in an asymmetric double well potential mimicking the H-bond**. The deepest well on the left corresponds to the energy minimum of the proton bounded to an oxygen by covalent bond at distance d_{OH}; the barrier is due to the energy that must be spent in order to break the covalent bond and make the proton hop to the second oxygen. (a) The effect of Zero Point Energy (ZPE) is to keep the proton above the Potential Energy Surface (PES), even in the ground state, making easier to overcome the barrier by thermal effects and changing the way potential energy landscape is explored. For the same reason, the proton is delocalized. (b) Proton tunneling allow the proton to overcome the potential energy barrier and then switch the covalent bond with the hydrogen one (in this case the asymmetric double well potential would be inverted).

3.3.2 Path integral simulations of water

There are several methods to simulate quantum nuclei, one of the most common being the path integral approach, which is based on Feynman's path integral formulation of quantum mechanics [153]. This approach leverages the isomorphism between quantum mechanics and classical statistical mechanics of polyatomic fluids.

This isomorphism was first exploited in Path integral molecular dynamics (PIMD) [154] and Path integral Monte Carlo (PIMC) [155] simulations in the early 1980s. Since then, advances in computational power have made these methods more mainstream [156].

Water systems were among the first subjects studied using PIMC [157] and PIMD [158] simulations based on empirical force fields. More recent simulations of bulk water demonstrate that NQEs influence macroscopic properties even at room temperature [32, 159].

Protonated water has also been extensively studied using these methods, in combination with *ab initio* techniques, both in clusters [160, 161] and bulk water [48, 162]. PIMD simulations have shown that the hydrated proton forms a fluxional defect in the H-bond network, rather than existing in a specific hydration state. These simulations also revealed that the small potential barrier is essentially washed out by ZPE, making proton delocalization important, while proton tunneling is negligible [48]. This is in contrast to what has been observed in simulations of the ice-X phase [163] and measured in small neutral water clusters [164, 165], as well

as larger clusters [166].

3.3.3 From quantum path integrals to classical ring polymers

Following Ref. [167], our starting point is the Hamiltonian in Eq. (1.3). Using again

$$H(\mathbf{p},\mathbf{q}) = \sum_{a=1}^{N} \frac{\mathbf{p}_a^2}{2m_a} + H_{\mathrm{e}}(\mathbf{q}) = \hat{T}_{\mathrm{n}}(\mathbf{p}) + \hat{V}_{\mathrm{n}}(\mathbf{q}), \tag{3.59}$$

where $\hat{V}_{\mathrm{n}}(\mathbf{q})$ is the potential operator. Its value obtained within the BO approximation is the PES $E_0(\mathbf{q})$, which can be numerically evaluated by means of the methods described in the previous Chapter. In what follows, for the sake of readability, we will drop the n subscript from $\hat{T}_{\mathrm{n}}$ and $\hat{V}_{\mathrm{n}}$, since we will deal only with the nuclear degrees of freedom.

In classical statistical mechanics the given ensemble implies a certain probability density function defined on the phase space, $\rho(\Gamma)$, which is used to compute average values of observables. In quantum statistical mechanics this role is played by the *density matrix*, which in the case of the Hamiltonian eigenstates defined in 1.14, $\Omega_k(\mathbf{q})$, can be written as

$$\rho = \sum_k |\Omega_k\rangle \frac{\langle\Omega_k| e^{-\beta H} |\Omega_k\rangle}{\mathcal{Z}} \langle\Omega_k| = \sum_k f(E_k) |\Omega_k\rangle \langle\Omega_k|, \tag{3.60}$$

where we have defined the coefficients $f(E_k)$ necessary to describe the canonical ensemble,

$$f(E_k) = \frac{e^{-\beta E_k}}{\mathcal{Z}}, \tag{3.61}$$

where the normalization constant $\mathcal{Z}$ is the quantum partition function, defined as

$$\mathcal{Z}(N,V,T) = \mathrm{Tr}\left[e^{-\beta H}\right], \tag{3.62}$$

over which any observable can be averaged, such that

$$\langle A\rangle = \frac{1}{\mathcal{Z}} \mathrm{Tr}\left[A e^{-\beta H}\right]. \tag{3.63}$$

Partition functions of quantum particle systems are less trivial than their classical counterparts because one has to take into account the Bose or Fermi statistics. However, we assume that the particles are distinguishable, while still treating them as quantum objects. Distinguishable quantum particles are also called Boltzmannons. Moreover, since the trace is basis-invariant, we will work in the position basis and express the partition function as an integral in $3N$-dimensions,

$$\mathcal{Z}(N,V,T) = \int \mathrm{d}\mathbf{q} \langle\mathbf{q}| e^{-\beta H} |\mathbf{q}\rangle, \tag{3.64}$$

where $\langle\mathbf{q}| e^{-\beta H} |\mathbf{q}\rangle$ are matrix diagonal elements. Since the expression in Eq. 3.64 cannot be solved analytically, one has to resort to approximations to evaluate $\mathcal{Z}$.

The first one is to consider small contributions coming from the exponential matrix,

$$e^{-\beta H} = \lim_{P\to\infty} \left(e^{-\beta H/P}\right)^P, \tag{3.65}$$

and evaluate each of the identical P small factors $e^{-\beta H/P}$ by inserting them in Eq. 3.64, sandwiched between $P-1$ resolutions of the identity:

$$I = \int \mathrm{d}\mathbf{q}\, |\mathbf{q}\rangle \langle \mathbf{q}| \,. \tag{3.66}$$

This results in

$$\begin{aligned} \mathcal{Z}(N,V,T) &= \int \mathrm{d}\mathbf{q}\, \mathrm{d}\mathbf{q}^2 \cdots \mathrm{d}\mathbf{q}^P \, \langle \mathbf{q}| e^{-\beta H} \left|\mathbf{q}^2\right\rangle \left\langle \mathbf{q}^2\right| e^{-\beta H} \left|\mathbf{q}^3\right\rangle \cdots \left\langle \mathbf{q}^{P-1}\right| e^{-\beta H} \left|\mathbf{q}^P\right\rangle \left\langle \mathbf{q}^P\right| e^{-\beta H} |\mathbf{q}\rangle = \\ &= \int \prod_{b=1}^{P} \mathrm{d}\mathbf{q}^b \left\langle \mathbf{q}^b\right| e^{-\beta H/P} \left|\mathbf{q}^{b+1}\right\rangle , \end{aligned} \tag{3.67}$$

where in the last passage we renamed $\mathbf{q} \to \mathbf{q}_1$ and we imposed the periodic boundary conditions by requiring that $\mathbf{q} = \mathbf{q}^{(1)} = \mathbf{q}^{(P+1)}$, to express the integral in a more compact form.

In order to evaluate the kinetic and potential contributions of the Hamiltonian we adopt a second approximation, namely the Trotter-Suzuki decomposition that we already employed before:

$$e^{-\beta H/P} \approx e^{-\beta \hat{V}/2P} e^{-\beta \hat{T}/P} e^{-\beta \hat{V}/2P}. \tag{3.68}$$

Since we are in the positions basis, it is easy to evaluate the potential energy term for each repeated factor, such as

$$\left\langle \mathbf{q}^b\right| e^{-\beta \hat{V}/2P} e^{-\beta \hat{T}/P} e^{-\beta \hat{V}/2P} \left|\mathbf{q}^{b+1}\right\rangle = e^{-\beta \hat{V}(\mathbf{q}^b)/2P} \left\langle \mathbf{q}^b\right| e^{-\beta \hat{T}/P} \left|\mathbf{q}^{b+1}\right\rangle e^{-\beta \hat{V}(\mathbf{q}^{b+1})/2P}. \tag{3.69}$$

For the kinetic energy term instead it is convenient to pass to the momentum representation,

$$\begin{aligned} \left\langle \mathbf{q}^b\right| e^{-\beta \hat{T}/P} \left|\mathbf{q}^{b+1}\right\rangle &= \int \mathrm{d}\mathbf{p}^b \, \mathrm{d}\mathbf{p}^{b+1} \left\langle \mathbf{q}^b|\mathbf{p}^b\right\rangle \left\langle \mathbf{p}^b\right| e^{-\beta \hat{T}/P} \left|\mathbf{p}^{b+1}\right\rangle \left\langle \mathbf{p}^{b+1}|\mathbf{q}^{b+1}\right\rangle = \\ &= \frac{1}{(2\pi\hbar)^{3N/2}} \int \mathrm{d}\mathbf{p}^b \, \mathrm{d}\mathbf{p}^{b+1} e^{i\mathbf{q}^b\cdot\mathbf{p}^b/\hbar} e^{-\frac{\beta[\mathbf{p}^b]^2}{2mP}} \underbrace{\left\langle \mathbf{p}^b|\mathbf{p}^{b+1}\right\rangle}_{\delta(\mathbf{p}^b-\mathbf{p}^{b+1})} e^{i\mathbf{q}^{b+1}\cdot\mathbf{p}^{b+1}/\hbar} = \\ &= \frac{1}{(2\pi\hbar)^{3N/2}} \int \mathrm{d}\mathbf{p}^b e^{i\mathbf{p}^b\cdot(\mathbf{q}^b-\mathbf{q}^{b+1})/\hbar} e^{-\frac{\beta[\mathbf{p}^b]^2}{2mP}}, \end{aligned} \tag{3.70}$$

and solve the resulting N-dimensional Gaussian integral:

$$\begin{aligned} \left\langle \mathbf{q}^b\right| e^{-\beta \hat{T}/P} \left|\mathbf{q}^{b+1}\right\rangle &= \frac{1}{(2\pi\hbar)^{3N/2}} \left(\frac{mP}{2\pi\beta}\right)^{3P/2} e^{\frac{mP}{2\beta\hbar^2}(\mathbf{q}^b-\mathbf{q}^{b+1})^2} \\ &= \left(\frac{mP}{2\pi\beta\hbar^2}\right)^{3N/2} e^{\frac{mP}{2\beta\hbar^2}(\mathbf{q}^b-\mathbf{q}^{b+1})^2}. \end{aligned} \tag{3.71}$$

By inserting (3.69) and (3.70) in (3.67) we get:

$$\mathcal{Z} = \lim_{P\to\infty} \left(\frac{mP}{2\pi\beta\hbar^2}\right)^{3P/2} \int \mathrm{d}\mathbf{q}^{(1)} \cdots \mathrm{d}\mathbf{q}^{(P)} \exp\left\{-\sum_{b=1}^{P}\left[\frac{mP}{2\beta\hbar}(\mathbf{q}^b-\mathbf{q}^{b+1})^2 + \frac{\beta}{P}V(\mathbf{q}^b)\right]\right\} \tag{3.72}$$

Notice that for finite P, which will forcely be the case of computer simulations, the single nuclear partition function is described as the *configurational integral* $\mathcal{Q}$ of a closed ring polymer made of

P beads indexed by the superscript b, with a nearest neighbour harmonic interaction and the external potential equivalent to the Born-Oppenheimer one, but scaled by β/P.

To sample the ring polymer phase space it is useful to apply the *quantum-classical mapping*, where, in addition to the beads positions, we also consider the fictitious momenta of the classical ring polymer. Then, the phase space element is given by

$$\mathbf{\Gamma}_{\mathrm{RP}} = (\mathbf{p}_1^1, \cdots \mathbf{p}_1^P, \mathbf{p}_2^1, \cdots, \mathbf{p}_{N-1}^P, \mathbf{p}_N^1, \cdots, \mathbf{p}_N^P, \mathbf{q}_1^1, \cdots \mathbf{q}_1^P, \mathbf{q}_2^1, \cdots, \mathbf{q}_{N-1}^P, \mathbf{q}_N^1, \cdots, \mathbf{q}_N^P), \tag{3.73}$$

and the canonical partition function at temperature β/P is the one of a system of N ring polymers of P beads each interacting via an harmonic potentials between the nearest neighbours of a same necklace, and subjected to and external potential which is given by the electrons:

$$\mathcal{Z}_{\mathrm{RP}} \propto \int \mathrm{d}\mathbf{q}^1 \cdots \mathrm{d}\mathbf{q}^P \, \mathrm{d}\mathbf{p}^1 \cdots \mathrm{d}\mathbf{p}^P \exp\left\{ -\sum_{b=1}^{P} \beta \left[\frac{[\mathbf{p}^b]^2}{2\mu} + \frac{mP}{2\beta^2\hbar}(\mathbf{q}^b - \mathbf{q}^{b+1})^2 + \frac{\beta}{P} V(\mathbf{q}^b) \right] \right\}. \tag{3.74}$$

The classical Hamiltonian of the ring polymer is

$$H_{\mathrm{RP}} = \sum_{b=1}^{P} \left[\sum_{a=1}^{N} \left(\frac{[\mathbf{p}_a^b]^2}{2\mu_a^b} + \frac{1}{2} m\omega^2 (\mathbf{q}_a^b - \mathbf{q}_a^{b+1})^2 \right) + \frac{1}{P} V(\mathbf{q}^b) \right], \tag{3.75}$$

where we used fictitious masses μ and we expressed the harmonic constant as

$$\omega = \frac{\sqrt{P}}{\beta\hbar}. \tag{3.76}$$

This Hamiltonian allows one to propagate the beads of the ring polymer according to the usual equation of motion:

$$\begin{cases} \dot{\mathbf{p}}_a^b = -m_a\omega^2(2\mathbf{q}_a^b - \mathbf{q}_a^{b-1} - \mathbf{q}_a^{b+1}) + \frac{1}{P}\nabla_{\mathbf{q}_a^b} V \\ \dot{\mathbf{q}}_a^b = \mathbf{p}_a^b/\mu_a \end{cases} \tag{3.77}$$

a technique named Ring polymer molecular dynamics (RPMD).

These equations are not easy to integrate because of the slow convergence of Eq. (3.77): the harmonic term increases with the number of beads P, resulting in stiffer vibration modes, while the potential term V decreases with P, i.e. the molecular vibrations due to the BO PES become less important [168]. To solve this issue we switch to *normal modes* coordinates, so that the harmonic oscillators are decoupled. This is equivalent to diagonalize the matrix $\mathbf{M}$ used to define the quadratic form of the harmonic interactions:

$$M^{b,c} = 2\delta^{b,c} - \delta^{b,c+1} - \delta^{b,c-1} \qquad b, c \in [1, \cdots, P], \tag{3.78}$$

where b and c refers to beads of the same ring polymer, and the matrix satisfies periodic boundary condition in the rows. The transformation is found simply by constructing the unitary matrix $\mathbf{U}$ of eigenvectors of $\mathbf{M}$:

$$\tilde{\mathbf{q}}_a^b = \frac{1}{\sqrt{P}} \sum_{c=1}^{P} U_{bc} \mathbf{q}_a^c. \tag{3.79}$$

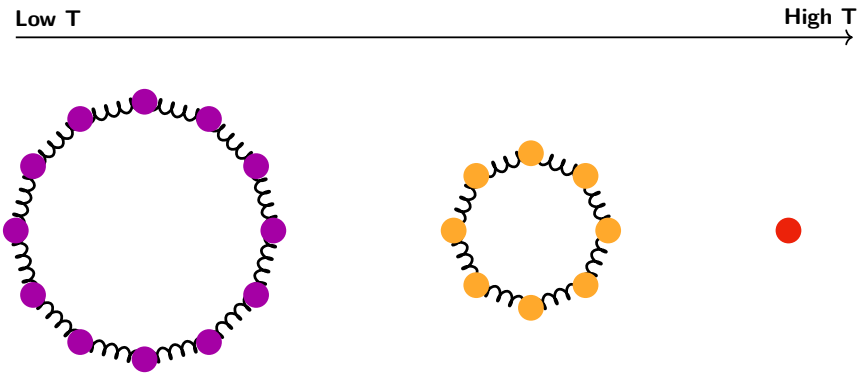


Figure 3.2: **Quantum-classical ring polymer isomorphism.** Intuitive picture of the quantum-classical isomorphism: at low temperature light nuclei are described by ring polymers made of many replicas of the particles interacting via an harmonic potential to simulate their true quantum nature.

This allows one to rewrite the ring polymer Hamiltonian with decoupled harmonic oscillators:

$$\widetilde{\widehat{H}}_{\mathrm{RP}} = \sum_{b=1}^{P}\sum_{a=1}^{N}\left[\frac{[\mathbf{p}_a^b]^2}{2m_a^b} + \frac{1}{2}m\omega_P^2\lambda_b\left[\tilde{\mathbf{q}}_a^b\right]^2\right] + \frac{1}{P}\sum_{b=1}^{P}V(\mathbf{q}^b(\tilde{\mathbf{q}}_1^b,\cdots,\tilde{\mathbf{q}}_N^b)), \tag{3.80}$$

with λ's eigenvalues of $\mathbf{M}$,

$$\lambda_{2k-1} = \lambda_{2k-2} = 2P\left[1-\cos\left(\frac{2\pi(k-1)}{P}\right)\right]. \tag{3.81}$$

3.4 Ring polymer molecular dynamics at zero temperature

Time evolution of the ring polymer can be computed by an algorithm analogous to the velocity-Verlet presented in Section 3.1.3, with the only difference that a back and forth normal modes transformation needs to be applied in the position propagation step.

We still use symmetric splitting of the propagator:

$$e^{i\hat{L}^{\mathrm{RP}}\delta t} \approx e^{i\hat{L}_{\mathbf{p}}^{\mathrm{RP}}\frac{\delta t}{2}}e^{i\hat{L}_{\mathbf{q}}^{\mathrm{RP}}\delta t}e^{i\hat{L}_{\mathbf{p}}^{\mathrm{RP}}\frac{\delta t}{2}}, \tag{3.82}$$

where we used the superscript RP to indicate that we are propagating the whole ring polymer.

1. Propagate the particle momenta after $\delta t/2$ using the forces derived from the potential of the RP Hamiltonian:

$$\mathbf{p}_a^b(t+\delta t/2) = e^{i\hat{L}_{\mathbf{p}}\delta t/2}\mathbf{p}_a^b(t) = \left(1+\frac{\delta t}{2}\mathbf{f}_a^b\cdot\nabla_{\mathbf{p}_a^b}\right)\mathbf{p}_a^b(t) = \mathbf{p}_a^b(t)+\frac{\delta t}{2}\mathbf{f}_a^b(t); \tag{3.83}$$

2. Switch to normal mode coordinates and propagate them of a time step δt according to the free RP Hamiltonian Eq. (3.75), then switch back again to cartesian coordinates:

$$\begin{aligned}\tilde{\mathbf{p}}_a^b &\leftarrow \frac{1}{\sqrt{P}}\sum_{c=1}^{P}U_{bc}\mathbf{p}_a^c,\\ \tilde{\mathbf{q}}_a^b &\leftarrow \frac{1}{\sqrt{P}}\sum_{c=1}^{P}U_{bc}\mathbf{q}_a^c,\end{aligned} \tag{3.84}$$

$$\begin{pmatrix} \tilde{\mathbf{p}}_a^b(t+\delta t) \\ \tilde{\mathbf{q}}_a^b(t+\delta t) \end{pmatrix} = \begin{pmatrix} \cos\omega\delta t & -m_a\omega\sin\omega\delta t \\ \frac{1}{m_a\omega}\sin\omega\delta t & \cos\omega\delta t \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{p}}_a^b(t) \\ \tilde{\mathbf{q}}_a^b(t) \end{pmatrix}; \tag{3.85}$$

$$\begin{aligned} \mathbf{p}_a^b &\leftarrow \frac{1}{\sqrt{P}}\sum_{c=1}^{P} U_{bc}^{\dagger}\tilde{\mathbf{p}}_a^c, \\ \mathbf{q}_a^b &\leftarrow \frac{1}{\sqrt{P}}\sum_{c=1}^{P} U_{bc}^{\dagger}\tilde{\mathbf{q}}_a^c, \end{aligned} \tag{3.86}$$

3. Switch back from normal mode to real coordinates in order to evaluate the forces in the new positions

$$\mathbf{f}_a^b(t+\delta t) = -\nabla_{\mathbf{q}^b} E_0\left(\mathbf{q}^b(t+\delta t)\right). \tag{3.87}$$

4. Propagate the particle momenta for the remaining half time step $\delta t/2$

$$\mathbf{p}_a^b(t+\delta t) = \mathbf{p}_a^b(t+\delta t/2) + \frac{\delta t}{2}\mathbf{f}_a^b(t+\delta t/2); \tag{3.88}$$

3.5 Ring polymer molecular dynamics at finite temperature

As for the classical counterpart, there are different schemes for finite temperature simulations of a quantum system, from the deterministic Nosé-Hoover chain [169] to stochastic thermostatting algorithms, the latter complessively designated as Path integral Langevin dynamics (PILD). In *normal modes* representation, where the bead momenta $\{\mathbf{p}^b\}_{b=1,\cdots,P}$ are rotated into $\tilde{\mathbf{p}}^k$, the corresponding under-damped Langevin equation are:

$$\begin{cases} \dot{\tilde{\mathbf{p}}}^k = -m\omega_b^2\tilde{\mathbf{q}}^k - \overline{\gamma}^k\mathbf{p}^k + \boldsymbol{\eta}^k(t) \\ \dot{\tilde{\mathbf{q}}}^k = \frac{\tilde{\mathbf{q}}}{\mu}, \end{cases} \tag{3.89}$$

where the noise vector is still defined by Gaussian white noise $\boldsymbol{\xi}^k$ multiplied by a factor which accounts not only the friction matrix $\overline{\gamma}^k$, but also for the ring polymer temperature and the number of beads:

$$\boldsymbol{\eta}^k(t) = \sqrt{\frac{2m\overline{\gamma}^k P}{\beta}}\boldsymbol{\xi}_t^k, \tag{3.90}$$

and $\omega_k = 2\tilde{\omega}_P \sin\frac{(k-1)\pi}{P}$ is the frequency of the k-th harmonic mode.

Two examples of Path integral Langevin integrators are the Path integral Langevin equation (PILE) [170], which is the quantum version of the Bussi algorithm, and the Path integral Ornstein-Uhlenbeck process (PIOUD) [133] algorithm.

3.5.1 Path Integral Langevin Equation

The Trotter-Suzuki breakup in the Path integral Langevin equation (PILE) algorithm is analogous to Eq. (3.31),

$$e^{i\hat{L}_{\mathrm{PILE}}\delta t} = e^{i\hat{L}_{\mathrm{FP}}\frac{\delta t}{2}} \underbrace{e^{i\hat{L}_{\mathbf{p}}^{\mathrm{RP}}\frac{\delta t}{2}} e^{i\hat{L}_{\mathbf{q}}^{\mathrm{RP}}\delta t} e^{i\hat{L}_{\mathbf{p}}^{\mathrm{RP}}\frac{\delta t}{2}}}_{e^{i\hat{L}_{\mathrm{RP}}\delta t}} e^{i\hat{L}_{\mathrm{FP}}\frac{\delta t}{2}}, \tag{3.91}$$

where in place of the velocity-Verlet propagation there is the procedure described in Section 3.4. More in details, the steps are the following:

1. Switch from real to normal mode coordinates to apply the exact thermostatting.

$$\tilde{\mathbf{p}}^k(t^+) = e^{i\hat{L}_{\mathrm{FP}}\frac{\delta t}{2}}\tilde{\mathbf{p}}^k = c_1^k\tilde{\mathbf{p}}^k(t) + \sqrt{\frac{mP}{\beta}}c_2^k\tilde{\boldsymbol{\xi}}^k. \tag{3.92}$$

2. Propagate the ring polymer according to the RP Hamiltonian(3.75), following the three steps described in Section 3.4.

3. Repeat step one to finally thermostatting the last half time step $\delta t/2$.

$$\tilde{\mathbf{p}}^k(t+\delta t) = e^{i\hat{L}_{\mathrm{FP}}\frac{\delta t}{2}}\tilde{\mathbf{p}}^k(t^- + \delta t/2) = c_1^k\tilde{\mathbf{p}}^k(t^- + \delta t/2) + \sqrt{\frac{mP}{\beta}}c_2^k\tilde{\boldsymbol{\xi}}^k. \tag{3.93}$$

In the above steps, the values of c_1^b and c_2^b are specified as those in the Bussi algorithm (Eq. 3.37):

$$c_1^k = e^{-\overline{\gamma}^k\frac{\delta t}{2}}c_2^k = \sqrt{1-\left[c_1^k\right]}. \tag{3.94}$$

In the normal mode representation, the optimal choice of $\overline{\gamma}^k$ is [170]

$$\overline{\gamma}^k = \begin{cases} 1/\tau_0 & k = 0 \\ 2\omega_k & k > 0 \end{cases} \tag{3.95}$$

where τ_0 is a separate thermostat time constant for the centroid.

3.5.2 Path integral Ornestein-Uhlenbeck dynamics

The quantum generalization of the CMPC algorithm, namely the Path Integral Momentum-Position correlator, is based on a matrix similar to the one in (3.53), but of size $6NP \times 6NP$ to account for the interbeads harmonic forces:

$$\hat{\overline{\gamma}} = \begin{pmatrix} \overline{\gamma} & \mathbf{K} \\ -\mathbf{I} & \mathbf{0} \end{pmatrix}, \tag{3.96}$$

where $\mathbf{K}$ is a $3NP \times 3NP$ matrix defined as

$$K_{b_1a_1x_1,b_2a_2x_2} = \tilde{\omega}_P^2\delta_{x_1x_2}\delta_{a_1a_2}\left(2\delta^{b_1,b_2} - \delta^{b_1,b_2-1} - \delta^{b_1,b_2+1}\right), \tag{3.97}$$

where the row and column indices, separated by the comma, are described as a collection of indices for the beads b, which must be contiguous, and the atom a and cartesian coordinates x, which must be the same, because the harmonic forces couples same atom components.

Unfortunetaly in this case, the simultaneous propagation of positions and momenta would lead to a sub-optimal sampling, because soft modes of molecular vibration are overdamped by the ring polymer vibration modes, which become stiffer as P increases.

The Path integral Ornstein-Uhlenbeck process (PIOUD) algorithm represents both an improvement and generalization to stochastic forces of the methods illustrated before. It relies on a different decomposition of the Fokker-Planck propagator, based on the separation of physical modes and the fictitious harmonic modes [133],

$$i\hat{L}_{\mathrm{KK}} = \sum_{i=1}^{3NP} \underbrace{\left[\mathcal{F}_i \nabla_{\mathbf{p}_i} + \mathbf{p}_i \nabla_{\mathbf{q}_i}\right]}_{\text{Hamiltonian}} - \sum_{j=1}^{3NP} \underbrace{\gamma_{ij}\left(\nabla_{\mathbf{p}_i}\mathbf{p}_j + k_B T P \nabla_{\mathbf{p}_i}\nabla_{\mathbf{p}_j}\right)\Big]}_{\text{thermostat}} \tag{3.98}$$

where the indices run over all particles and replicas.

To solve this problem, the two contribution can be treated separately, by decomposition of the generalized force $\mathcal{F}$ and the friction matrix in Born-Oppenheimer and harmonic oscillator terms:

$$\mathcal{F} = \mathbf{f}^{\mathrm{BO}} + \mathbf{f}^{\omega}, \tag{3.99}$$

$$\overline{\gamma} = \overline{\gamma}^{\mathrm{BO}} + \overline{\gamma}^{\omega}. \tag{3.100}$$

Consequently, also the stochastic Liouvillian can be decomposed in:

$$i\hat{L}_{\mathrm{KK}} = i\hat{L}_{\mathrm{BO}} + i\hat{L}_{\omega} \tag{3.101}$$

$$i\hat{L}_{\omega} = \sum_{i=1}^{3NP}\left[\mathbf{f}_i^{\omega}\nabla_{\mathbf{p}_i} + \mathbf{p}_i\nabla_{\mathbf{q}_i} - \sum_{j=1}^{3NP}\gamma_{ij}^{\omega}\left(\nabla_{\mathbf{p}_i}\mathbf{p}_j + k_B T P \nabla_{\mathbf{p}_i}\nabla_{\mathbf{p}_j}\right)\right] \tag{3.102}$$

$$i\hat{L}_{\mathrm{BO}} = \sum_{i=1}^{3NP}\left[\mathbf{f}_i^{\mathrm{BO}}\nabla_{\mathbf{p}_i} + \mathbf{p}_i\nabla_{\mathbf{q}_i} - \sum_{j=1}^{3NP}\gamma_{ij}^{\mathrm{BO}}\left(\nabla_{\mathbf{p}_i}\mathbf{p}_j + k_B T P \nabla_{\mathbf{p}_i}\nabla_{\mathbf{p}_j}\right)\right] \tag{3.103}$$

$$e^{i\hat{L}_{\mathrm{KK}}\delta t} \approx e^{i\hat{L}_{\mathrm{BO}}\delta t/2} e^{i\hat{L}_{\omega}\delta t} e^{i\hat{L}_{\mathrm{BO}}\delta t/2} \tag{3.104}$$

Steps:

1. Update the particles momenta according on the knowledge of the BO- and stochastic forces. This translates onto using an equation equivalent to the general solution of the Langevin equation as Eq. (3.42), with the assumption that $\mathbf{f}^{\mathrm{BO}}$ and $\overline{\gamma}^{\mathrm{BO}}$ are constant in the small timestep, as in Eq. (3.43):

$$\mathbf{p}(t^- + \delta t/2) = \mathbf{p}(t) e^{-\overline{\gamma}^{\mathrm{BO}}\delta t/2} + \int_t^{t+\delta t/2} \mathrm{d}s e^{\overline{\gamma}_{\mathbf{q}}(t'-s)}\left[\mathbf{f}_{\mathbf{q}}^{\mathrm{BO}} + \boldsymbol{\eta}(s)\right] \tag{3.105}$$

   In case of deterministic forces, the update can be done in the real coordinates space, while for stochastic forces first one must switch to the frame that diagonalizes the Langevin damping matrix $\overline{\gamma}^{\mathrm{BO}}$.

2. Apply a back and forth normal mode transformation that propagates the harmonic part by δt and thermalizes the ring polymer, according to the equations:

$$\mathbf{p}(t^+ + \delta t/2) = \boldsymbol{\Lambda}_{1,1}\mathbf{p}(t^- + \delta t/2) + \boldsymbol{\Lambda}_{1,2}\mathbf{q}(t) + \boldsymbol{\Gamma}\tilde{\boldsymbol{\eta}} \tag{3.106}$$

$$\mathbf{q}(t + \delta t) = \boldsymbol{\Lambda}_{2,1}\mathbf{p}(t^- + \delta t/2) + \boldsymbol{\Lambda}_{2,2}\mathbf{q}(t) + \boldsymbol{\Theta}\tilde{\tilde{\boldsymbol{\eta}}} \tag{3.107}$$

where the only forces contribution comes from the harmonic couplings, and not the BO-forces. Notice that the update of the position is δt, while for the momenta we complete the half step in BO- and stochastic fores started at point 1, while the harmonic forces are fully propagated for the entire time step. Since we are missing the last half BO and stochastic contribution, we use the notation $t^+ + \delta/2$. The matrices $\mathbf{\Lambda}$ are the lengthy results of analytic integration, reported in Ref. [133]. The noise and forces $6NP$-dimensional contribution are computed as:

$$\mathbf{\Xi} = \begin{pmatrix} \mathbf{\Gamma}\tilde{\boldsymbol{\eta}} \\ \mathbf{\Theta}\tilde{\tilde{\boldsymbol{\eta}}} \end{pmatrix} = \int_{t_n}^{t_{n+1}} \mathrm{d}t e^{\hat{\tilde{\gamma}}(t-t_{n+1})} \begin{pmatrix} \boldsymbol{\eta} \\ \mathbf{0} \end{pmatrix}, \tag{3.108}$$

$$\mathbf{F} = \begin{pmatrix} \mathbf{\Gamma}\mathbf{f}_n \\ \mathbf{\Theta}\mathbf{f}_n \end{pmatrix} = \hat{\tilde{\gamma}}^{-1} \left(\mathbf{I} - e^{\hat{\tilde{\gamma}}(t-t_{n+1})} \right) \begin{pmatrix} \mathbf{f}_n \\ \mathbf{0} \end{pmatrix}, \tag{3.109}$$

3. Evaluate the ionic forces in the new positions via

$$\mathbf{f}(t+\delta t) = -\nabla_{\mathbf{q}} V\left(\mathbf{q}(t+\delta t)\right) \tag{3.110}$$

4. Update again the particle momenta for the last half of time step, $\delta t/2$, as done in step 1.

$$\mathbf{p}(t+\delta t) = \mathbf{p}(t^+ + \delta/2) e^{-\overline{\gamma}^{\mathrm{BO}} \delta t/2} + \int_t^{t+\delta t/2} \mathrm{d}s e^{\overline{\gamma}_{\mathbf{q}}(t'-s)} \left[\mathbf{f}_{\mathbf{q}}^{\mathrm{BO}} + \boldsymbol{\eta}(s) \right] \tag{3.111}$$

PIOUD can be used also with deterministic forces. In that case, $\gamma_{\mathrm{BO}} = 0$, as there is no need of correcting BO forces with an additional Langevin thermostat, since they are not affected by any noisy contribution. We would like to note that in case of deterministic forces it is always more convenient to use PIOUD rather than PILE, because there is one less Trotter breakup in the former integrator. Indeed, in PIOUD the Liouvillian factor related to $e^{iL_{\mathrm{FP}}\delta t/2} e^{iL_{\mathbf{q}}^{\mathrm{RP}}\delta t} e^{iL_{\mathrm{FP}}\delta t/2}$ is integrated in a single shot, without breaking it into three factors as in PILE. This feature allows one to use larger time steps in PIOUD for an enhanced stability.

3.5.3 Ring polymer and QMC: bead-grouping approximation

Usually, *ab initio* RPMD studies are based on a PES provided by DFT, for which the computational cost of force evaluation is necessarily proportional to P. Therefore, most of the techniques proposed to lighten this computational burden focus on decreasing the number of evaluations of the ionic forces. This has been achieved by ring polymer contraction [171, 172], or by reducing the number of quantum replicas using generalized Langevin Equations that leverage colored noise that mimics nuclear quantum fluctuations [173].

Although these methods could be effectively incorporated in a QMC framework too, the main computational bottleneck in our case is the large number of variational parameters, rather than the large value of P. Indeed, each bead at each iteration has its own optimal wavefunction, $|\Psi_{\mathbf{q}}^{(k)}\rangle$, for $k = 1, \dots, P$, which minimizes the variational energy at the nuclear configuration $\mathbf{q}^b$. Consequently, we need to find the best variational parameters set,

$$\boldsymbol{\lambda}^{(k)} = \{ g_{\mu,\nu}^{a^k,b^k}, \lambda_{\mu,\nu}^{a^k,b^k}, b^k, \zeta_{l,n}^k, \dots \}, \tag{3.112}$$

for each wavefunction.

To overcome this major difficulty, we exploit the *local* nature of the Gaussian basis sets used in the expansion of both the Jastrow and AGP factors. In fact, the most relevant dependence of the wavefunction on the ionic positions $\mathbf{q}$ comes explicitly from the basis set, and less from the electronic variational parameters, which depend on them only indirectly.

It is therefore convenient to make the approximation of defining N_{groups} groups of neighboring beads and constraining the wavefunction parameters to be equal for all beads in the same group. Since a group shares the same parameters, the corresponding energy gradients are then averaged over the quantum replicas constituting the group. In this way, we improve the statistics by a factor of P/N_{groups}. We obtain less noisy parameters even though the resulting wavefunction is not exactly optimized for each quantum replica. This approximation is systematically improvable between two extremes: if one takes $N_{\text{groups}} = P$, the electronic result is exact, whereas $N_{\text{groups}} = 1$ constitutes the roughest approximation. In the latter case, one performs a fully quantum dynamics with almost the same statistics as the one with classical nuclei.

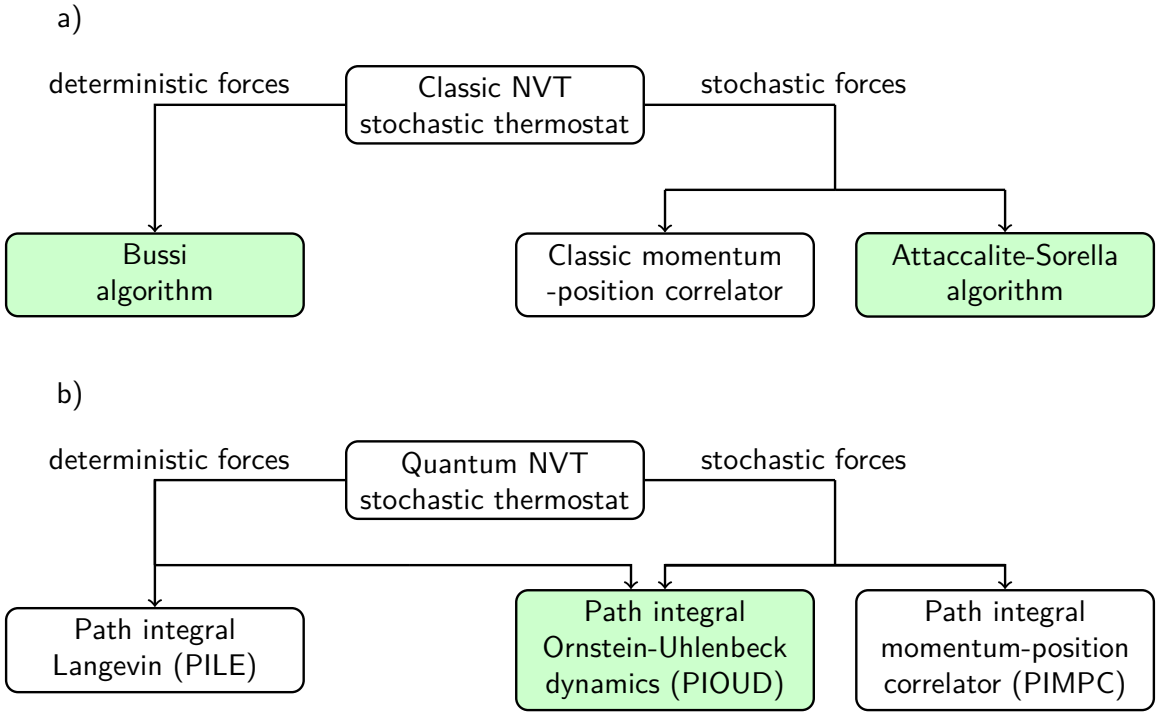


Figure 3.3: **Classical and quantum Langevin dynamics algorithms.** In green what it is used in this thesis

Chapter 4

Thermal dependence of the hydrated proton and optimal proton transfer in the protonated water hexamer

In this Chapter[1], we report our study of the protonated water hexamer $H_{13}O_6^+$ by MD simulations, fully retaining the nuclear quantum nature of the atoms using path integral methods (Chap. 3), and treating the electrons at the QMC level (Chap. 2). As we mentioned in the introduction, this system is the smallest protonated water cluster that includes the two limiting complexes involved in the proton transfer: the Eigen cation, which appears as $H_3O^+(H_2O)_5$, and the Zundel cation, included in the hexamer as $H_5O_2^+(H_2O)_4$. Both cations are fully solvated up to the first shell. Although the protonated water hexamer exhibits several isomers [174–176], in this work we consider its Zundel-like configuration because it is the one that most closely resembles the hydrated proton in bulk water, solvated up to the second shell (Figure 4.1). Even when the system will fall into distorted Eigen minimum, as it will described later, the structure of the second solvation shell will remain the one typically associated to the Zundel, with four water molecules.

To investigate the proton dynamics in the system, we start from the analysis of its potential energy surface, reported in Section 4.1, and compare it with the one of the Zundel cation as reported in Refs. [127, 133]. While the latter system misses a large part of water solvation effects, the former includes the full contribution of the first and second shells of the solvated proton.

In Section 4.2, we show that the Hydrogen bond (H-bond) mediated by the hydrated proton exhibits a remarkably low thermal expansion from zero temperature up to 300 K, with a nearly temperature-independent length that becomes *shorter* than the classical-ion counterpart in the [200-350] K temperature range. As we will see, the strength of the H-bond results from a non-trivial cooperation of NQEs and thermal activation. Indeed, NQEs strongly affect the vibrational levels of the proton shuttling mode bridging the central O_1 and O_2 oxygen atoms. These levels

[1]These results have been published in Ref. [127], on which most of this Chapter is based.

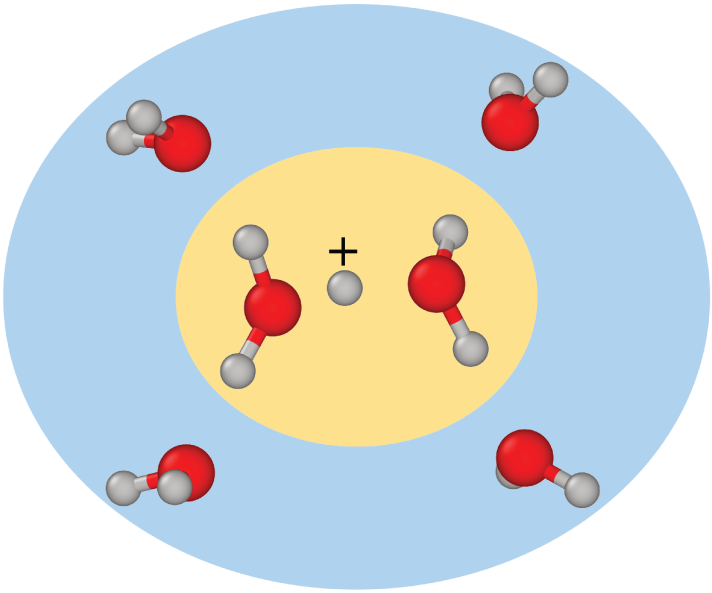


Figure 4.1: **Highlight of $H_{13}O_6^+$ in its Zundel configuration**, with a Zundel-core (yellow) solvated by four water molecules (light blue).

are then thermally occupied according to the $d_{O_1O_2}$ distance of a given configuration. We can thus distinguish three regimes (see Fig. 4.2):

(i) "short-Zundel" configurations with the shortest $d_{O_1O_2}$, where the proton along the shuttling mode feels a quadratic potential close enough to its energy minimum and it is perfectly shared between the two central water molecules;

(ii) "elongated-Zundel" configurations for intermediate $d_{O_1O_2}$, comprising the equilibrium distance, where a potential energy barrier starts to develop in between O_1 and O_2 and the proton is delocalised only due to NQEs;

(iii) "distorted-Eigen" configurations at even larger $d_{O_1O_2}$, where the central barrier is large enough that the hydrated proton is localised on one of the two flanking water molecules, forming an Eigen-like complex.

In Section 4.3 we will see that the occurrence of short-Zundel configurations is key to understand the H-bond thermal robustness and to enhance the proton transfer dynamics. Despite being energetically disfavoured by the short $d_{O_1O_2}$ distances at the classical level, these configurations are populated thanks to the synergistic action of NQEs and temperature, yielding a sweet spot for proton transfer in the [250-300] K temperature range. This last observation is supported by a 2-dimensional projection of the PES on the most relevant coordinates in Section 4.4. Once the thermal dependence of the structure of $H_{13}O_6^+$ is established, we can conduct the instanton dynamics analysis, which is presented in Section 4.5. The instanton is defined as a quantum proton configuration that connects *instantaneously* two steady states represented by

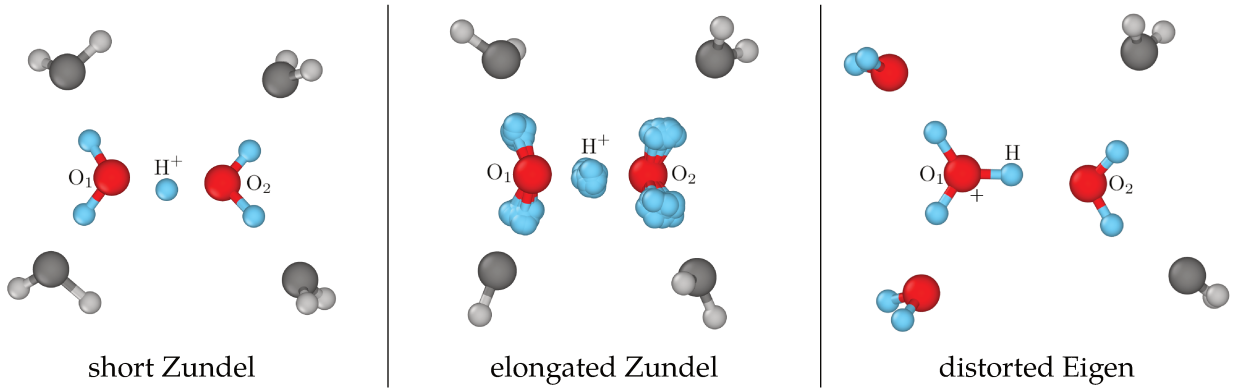


Figure 4.2: **Different regimes of the protonated water hexamer $H_{13}O_6^+$.** Left panel: short-Zundel configuration with a Zundel center ($H_5O_2^+$) in colors and its first solvation shell (4 H_2O) in gray shades. Central panel: elongated Zundel with the quantum nature of hydrogen atoms highlighted by the full representation of its imaginary-time positions in a PI configuration. Right panel: distorted-Eigen configuration with an Eigen cation ($H_9O_4^+$) in colors accompanied by two solvating water molecules (2 H_2O) in gray shades. The O_1, O_2 and H^+ labels are used throughout the paper to refer to the corresponding atoms, as indicated here.

the two minima of a double well potential, stretching across two water molecules. A summary of the results, motivating the work done in the second Part of this thesis, close this Chapter in the Discussion Section 4.6.

4.1 Role of solvation: Zundel ion versus protonated water hexamer

To quantify the impact of the solvation shell on the Zundel core, we compare in Fig. 4.3 the O_1-O_2 potential, $V_{O_1O_2}$ (left), and the corresponding classical equilibrium geometry (right) of the two clusters at various $d_{O_1O_2}$ (distance between the 2 central oxygen atoms). At short $d_{O_1O_2}$, the slope of the protonated hexamer $V_{O_1O_2}$ is slightly larger than the Zundel one, due to a greater electrostatic repulsion because of steric hindrance. At large $d_{O_1O_2}$, the protonated hexamer PES is softer than the Zundel one, because the solvating H_2O molecules enhance the polarisability of the core atoms. As we will see later, the balance between short- and long-range repulsion, once supplemented with the zero-point energie (ZPE), is key to quantify the relative abundance of short-Zundel and distorted-Eigen configurations, and thus, it allows for a quantitative understanding of the PT mechanism.

The VMC equilibrium O_1O_2 distance is found to be $d_{O_1O_2} = d_{\min} = 2.3930(5)$ Å, in good agreement with MP2 calculations for the protonated hexamer, the most widely used post Hartree-Fock theory to study water clusters. As mentioned in Chap. 2, VMC has a milder scale with the system size than MP2, allowing one to perform extensive calculations of the protonated hexamer.

We also find the $H_{13}O_6^+$ equilibrium $d_{O_1O_2}$, represented by a vertical dashed line in Fig. 4.3, to be ~ 0.05 Å, larger than the $H_5O_2^+$ one. More importantly, at variance with the Zundel cation, which is centrosymmetric [177, 178], the protonated water hexamer equilibrium geometry is

asymmetric with classical ions. This fundamental symmetry modification of the PES is induced by solvation effects, which tend to stabilize the hexamer into its elongated-Zundel configuration. This can rationalise some THz/FTIR absorption spectroscopy fingerprints of the solvated proton [179], which have been related to a fast inter-conversion between the (distorted-) Eigen and (short-) Zundel forms.

The PT energy barrier vanishes at $d_{O_1O_2} = d_{symm} \simeq 2.38$ Å, a distance that separates the short Zundel below from the elongated-Zundel configurations above. The height of the barrier is less than 100 K (in k_B units) at d_{min}, rapidly increasing as a function of $d_{O_1O_2}$. We therefore expect several consequences on the hydrated proton distribution and on its mobility at finite temperature, once the NQEs are taken into account.

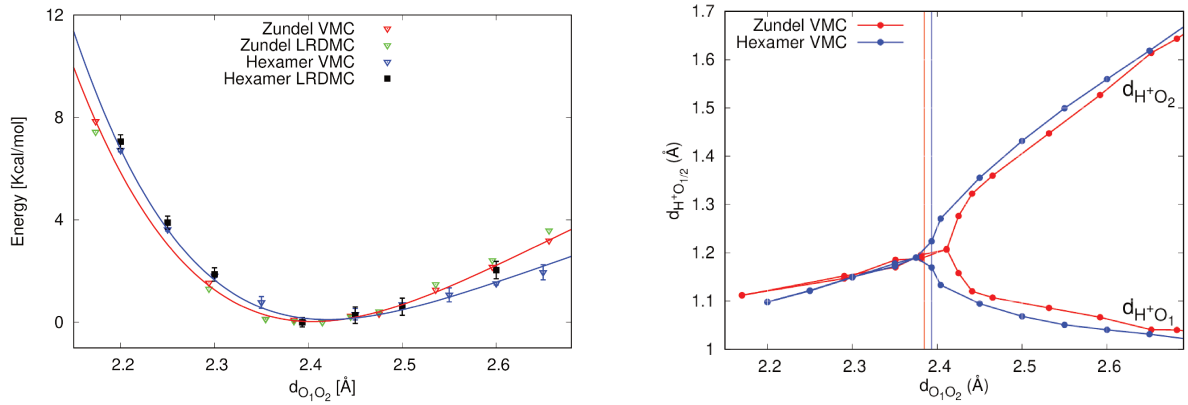


Figure 4.3: **Comparison of the protonated water dimer and hexamer $V_{O_1O_2}$ potential (left) and equilibrium geometry (right) as a function of $d_{O_1O_2}$.** Vertical dashed lines indicate the corresponding equilibrium $d_{O_1O_2}$. Notice that VMC and lattice regularised diffusion Monte Carlo (LRDMC) [180] energies are in nice statistical agreement for $d_{O_1O_2} \in$ [2.3-2.6] Å, the phase-space range explored by our MD simulations.

4.2 Thermal expansion of the H-bond

To understand how the dynamics of the hydrated proton evolves with temperature, QMC-driven AIMD simulations are relevant. Such calculations are carried out for both classical and quantum nuclei of the $H_{13}O_6^+$ ion, within the temperature interval $T \in$ [50-350] K, thanks to the methodological developments detailed in Ref. [133] and in the previous Chapters. At these conditions, the clusters are stable during the simulated time frame ($\approx$ 30 ps), allowing us to access the thermal properties of the hydrated proton and the $O_1H^+O_2$ bond over an extended temperature range.

From our QMC-MD simulations, we extract the normalised Pair Correlation Function (PCF) $g_{O_1O_2}$ for the two oxygen atoms O_1 and O_2 of the cluster core (Fig. 4.4). The expected broadening of the PCFs due to nuclear quantisation is significant over the whole temperature range (Fig. 4.4(b)). Only at temperatures as high as 350 K, the classical $g_{O_1O_2}$ (Fig. 4.4(a)) starts resembling the quantum distribution. This implies that the NQEs cannot be neglected for tem-

peratures up to this value, above ambient conditions. We also notice that, when comparing to the Zundel ion results [161], the peak position is shifted up by at least $\sim$ 0.01 Å. Thus, it appears that the $H_{13}O_6^+$ cluster frequently adopts elongated-Zundel configurations [50, 181, 182] at the lowest temperatures considered here. This is at variance with the protonated water dimer, where the hydrated proton lives in a single minimum symmetrically located between the two water molecules.

Focusing our attention to $\langle d_{O_1O_2} \rangle$ (Fig. 4.4(c)), its classical and quantum behaviours are remarkably different as a function of temperature. On the one hand, the classical $d_{O_1O_2}$ keeps increasing with temperature, as more energy is given to the intermolecular vibration modes. On the other hand, the quantum $d_{O_1O_2}$ displays a nearly flat behaviour with the cluster temperature, up to 300 K. This very low thermal expansion extended over a wide temperature range leads to a temperature regime where $d_{O_1O_2}$ for the quantum system become shorter than the classical values at the same temperatures. This is clearly seen in Fig. 4.4(c). We will come back to this point later.

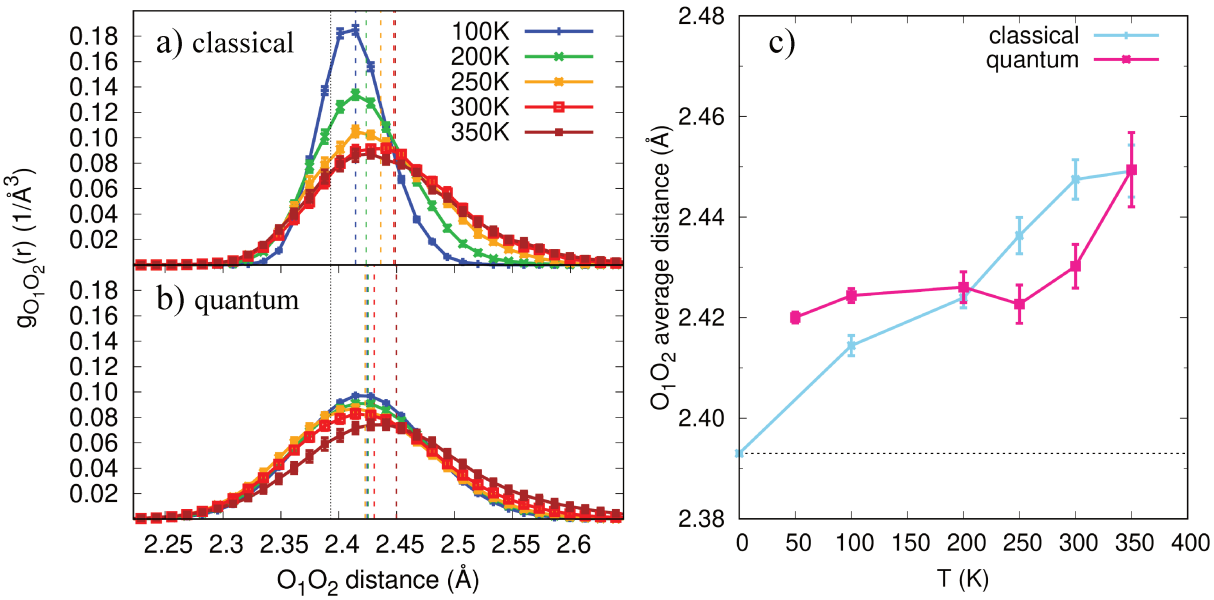


Figure 4.4: **Classical and quantum oxygen-oxygen $g_{O_1O_2}$ pair correlation functions as a function of temperature.** The dashed vertical lines indicate the average $\langle d_{O_1O_2} \rangle$ distance for each simulation, at the corresponding temperature. The dotted vertical line is located at the classical equilibrium geometry. Panel c) shows the T-dependence of the $\langle d_{O_1O_2} \rangle$ average distance. The classical equilibrium geometry is represented by a short-dashed horizontal black line. At 250 K and 300 K the oxygen-oxygen distance is *shortened* by NQEs with respect to the classical counterpart.

Finally, as the temperature further increases, the NQEs reduction weakens the central H-bond strength. Consequently, $d_{O_1O_2}$ spreads out, due to stochastic fluctuations of the core and the solvent, and a more classical regime is reached, when the averaged $d_{O_1O_2}$ values for classical and quantum nuclei meet again. The PCF distributions display longer tails, with more configu-

rations covering regions with $d_{O_1O_2} \in [2.5\text{-}2.7]$ Å, and the peak position rapidly shifts to larger values. Configurations with such a large $\langle d_{O_1O_2} \rangle$ are of distorted-Eigen type [181, 183].

4.3 A cooperative thermal-quantum species: the short-Zundel ion

To refine our structural analysis, we compute the bidimensional distribution function ρ_{2D}, which correlates the oxygen-oxygen (O_1O_2) and the oxygen-proton ($O_{1/2}H^+$, meaning that the oxygen can either be O_1 or O_2) distances, and study its temperature dependence $\rho_{2D} = \rho_{2D}(T)$. They are shown in Fig. 4.5 for both classical and quantum simulations.

To highlight the difference, in Fig. 4.6, we show the contour plot of the temperature-driven ρ_{2D} variation by taking $\rho_{2D}(250\,\text{K})$ as reference. Four temperature variations are explored: 100 K, 200 K, room temperature (RT), and 350 K (from the top to the bottom of Fig. 4.6).

In the *classical* protonated hexamer (Fig. 4.6, left column), rising the temperature from 250 K up to 350 K tends to stretch $\langle d_{O_1O_2} \rangle$, by promoting configurations from the elongated Zundel (blue central distribution with $d_{O_1O_2} \in [2.38\text{-}2.5]$ Å in Fig. 4.6) to an Eigen-like arrangement with larger $d_{O_1O_2}$ and a proton much more localised on one of the two central oxygen atoms (red wings). The situation is reversed at lower temperatures (100 K and 200 K) if compared to the 250 K reference, with positive (red) variations in the elongated Zundel and negative (blue) variations in the wings. Thus, for classical nuclei, there is a progressive depletion of the elongated Zundel and a corresponding population of the distorted-Eigen wings upon temperature rise. Short-Zundel configurations, highlighted in Fig. 4.6 by a gray background, seem to play a very marginal role in the temperature-driven density distribution shift.

The scenario is strikingly different with *quantum* nuclei (right column), particularly at the lowest temperatures (100 K and 200 K). In this regime, distorted-Eigen configurations are barely populated or depleted, and the density shift upon rising temperature takes place between the elongated-Zundel region and the short-Zundel sector. The latter is significantly more populated at 250 K than at lower temperatures at the expense of the elongated Zundel, which instead loses density with respect to the classical counterpart at the same temperature.

In the higher-temperature limit, at 350 K, NQEs are less relevant and, by consequence, the classical and quantum variations have a qualitatively similar behaviour. In both classical and quantum case, we notice the presence of red wings at large oxygen-oxygen distances ($d_{O_1O_2} \in [2.5\text{-}2.7]$ Å), which are the signature of thermally activated Eigen-like states, with a strongly localised proton. This is related to less frequent elongated-Zundel configurations, indicated by the depleted distribution for $d_{O_1O_2} < 2.5$ Å, confirming that the distorted-Eigen configurations are indeed promoted by high temperature. For quantum nuclei, the corresponding depletion goes well below the elongated-Zundel region, by touching also short-Zundel configurations, down to $d_{O_1O_2} \sim 2.3$ Å, at variance with the classical case, where the short-Zundel configurations are not involved.

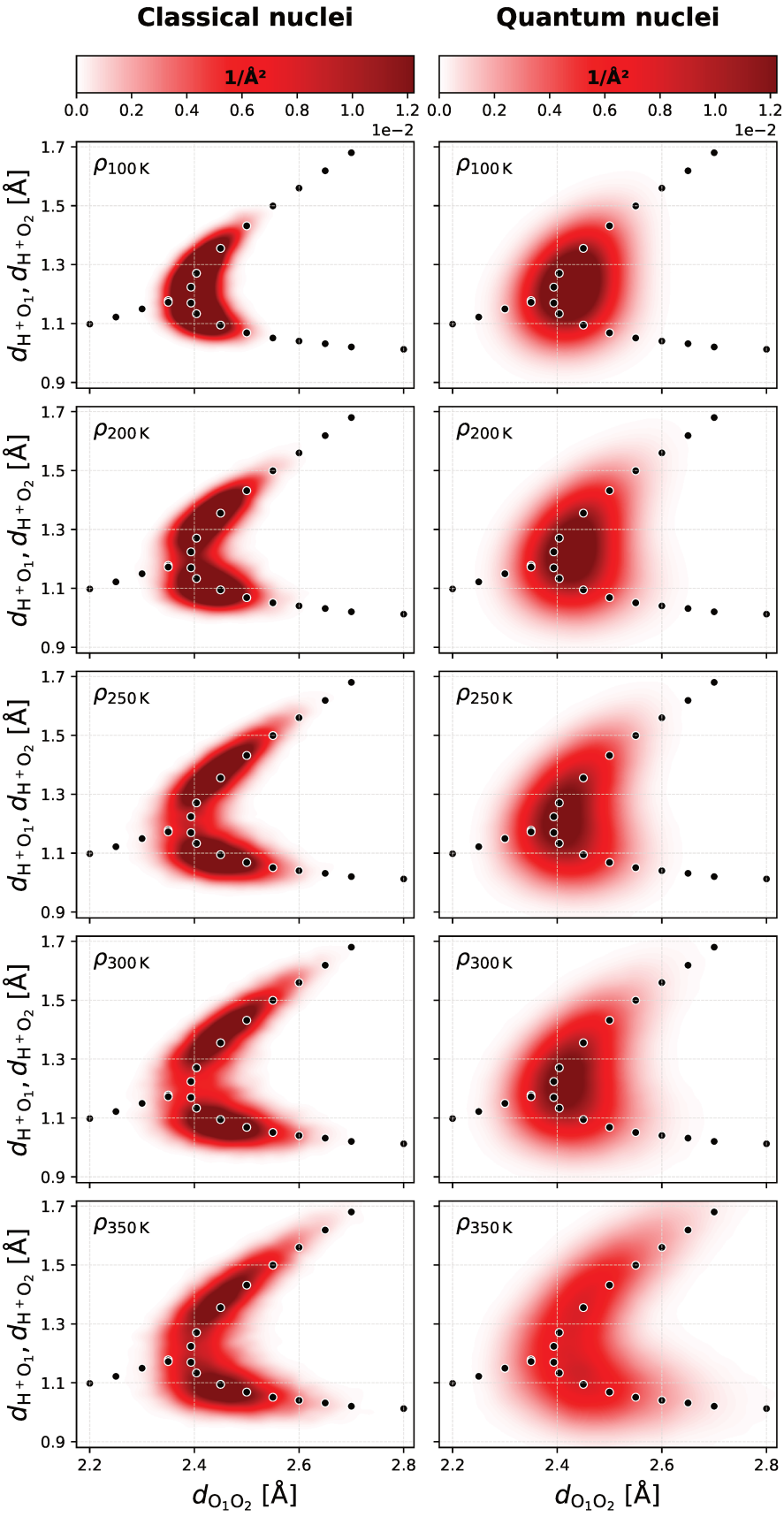


Figure 4.5: ρ_{2D} **computed from VMC-driven MD (left) and PIMD (right) at different temperatures.**

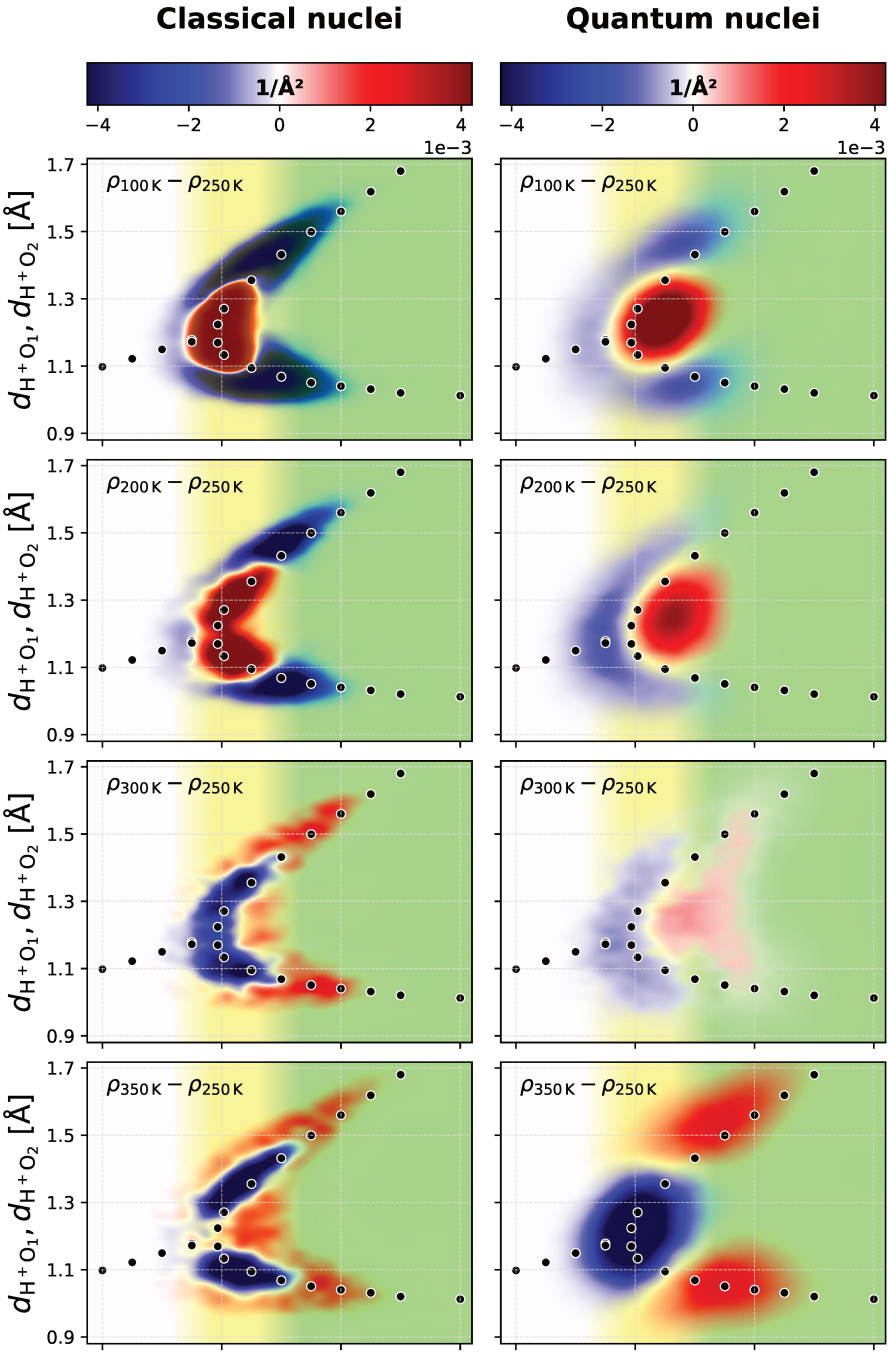


Figure 4.6: **Bidimensional oxygen-oxygen/oxygen-proton distributions.** Difference between bidimensional oxygen-oxygen/oxygen-proton distributions ρ_{2D} obtained by QMC-driven LD simulations for classical (left panels) and quantum (right panels) particles, computed at different temperatures. The bidimensional distribution computed at 250 K is taken as reference. Positive (negative) regions are in red (blue) color. The black filled circles correspond to the zero-temperature equilibrium geometries of the $H_{13}O_6^+$ ion at a fixed $d_{O_1O_2}$ distance. The coloured background highlights the three different regimes explained in the paper: the short Zundel (gray), the elongated Zundel (yellow), and the distorted Eigen (green) species.

4.4 Projected two-dimensional PES

To interpret these results, we first construct an accurate effective potential by projecting the full PES, computed during QMC-driven classical MD calculations, onto the degrees of freedom mostly relevant to understand the dynamics of the hydrated proton. These are the $d_{O_1O_2}$ distance and the proton sharing coordinate δ_{H^+}, referenced to the midpoint of the $O_1H^+O_2$ complex:

$$\delta_{H^+} \equiv d_{O_{1/2}H^+} - d_{O_1O_2}/2, \tag{4.1}$$

with $d_{O_{1/2}H^+}$ the $O_{1/2}$-H^+ distance projected onto the O_1O_2 direction. The resulting two-dimensional (2D) potential is $V_{2D} = V_{2D}(d_{O_1O_2}, \delta_{H^+})$. We refer the reader to Appendix B for technical details about the PES projection. We highlight that the potential V_{2D} is derived here at VMC quality. We also notice that δ_{H^+} is the vibrational coordinate of the proton shuttling mode, while $d_{O_1O_2}$ is related to the stretching mode of the two water molecules in the cluster core.

Given $V_{2D}(d_{O_1O_2}, \delta_{H^+})$, we then proceed to quantize the variable δ_{H^+}. Indeed, while $d_{O_1O_2}$ can be taken as classical, for it is related to the motion of heavier oxygen atoms of mass m_O, the δ_{H^+} coordinate must be quantised, owing to the light mass (m_H) of the hydrated proton. At the leading order in $2m_H/(m_O + m_H)$, we separate the stretching mode from the shuttling one, by invoking an adiabatic Born-Oppenheimer type of approximation (Chapter 1) for the two species [184]. We finally solve quantum-mechanically the Hamiltonian of a proton in the potential $V_{\delta_{H^+}} \equiv V_{2D}(\alpha, \delta_{H^+})|_{\alpha = d_{O_1O_2}}$ at fixed $d_{O_1O_2}$ value. In Fig. 4.7(a-c) we plot the ground state distribution and eigenvalues obtained for three distances, i.e. at $d_{O_1O_2} = 2.375$ Å, in the short-Zundel region close to the boundary between the short and the elongated Zundel, at $d_{O_1O_2} = 2.495$ Å, in the elongated-Zundel region close to the frontier between the elongated Zundel and the distorted Eigen, and finally at $d_{O_1O_2} = 2.585$ Å, deep into the distorted Eigen regime.

One can notice three different quantum behaviours of the vibrational shuttling mode, that provide a more quantitative ground to the three-regime distinction made at the beginning. In the short Zundel, $V_{\delta_{H^+}}$ is indeed a quadratic potential with a single minimum at the core center, which widens as $d_{O_1O_2}$ gets close to $d_{symm} \simeq 2.38$ Å, a distance where it becomes quartic because its curvature falls to zero before changing sign.

The ground state energy, i.e. the zero point energy (ZPE) of the shuttling mode, decreases as the potential widens, as reported in Fig. 4.7(d). In the elongated Zundel, a central barrier starts to develop, with a ground-state proton distribution that stays uni-modal thanks to a ZPE larger than its height, till $d_{O_1O_2} \simeq 2.5$ Å, where the ZPE equals the barrier height. In this regime, for $d_{O_1O_2} \in [d_{symm}, 2.5\,\text{Å}]$, the ZPE is particularly small, due to the quartic nature of $V_{\delta_{H^+}}$, and weakly $d_{O_1O_2}$-dependent, as shown in Fig. 4.7(d). Finally, for $d_{O_1O_2} > 2.5$ Å, we enter the distorted-Eigen regime, with an even larger central barrier > 1000 K, such that the quantum proton is instantaneously localised in one of the two wells, and its distribution is then bimodal. The ZPE starts to rise again as $d_{O_1O_2}$ is stretched, with a slope steeper - in absolute value - than

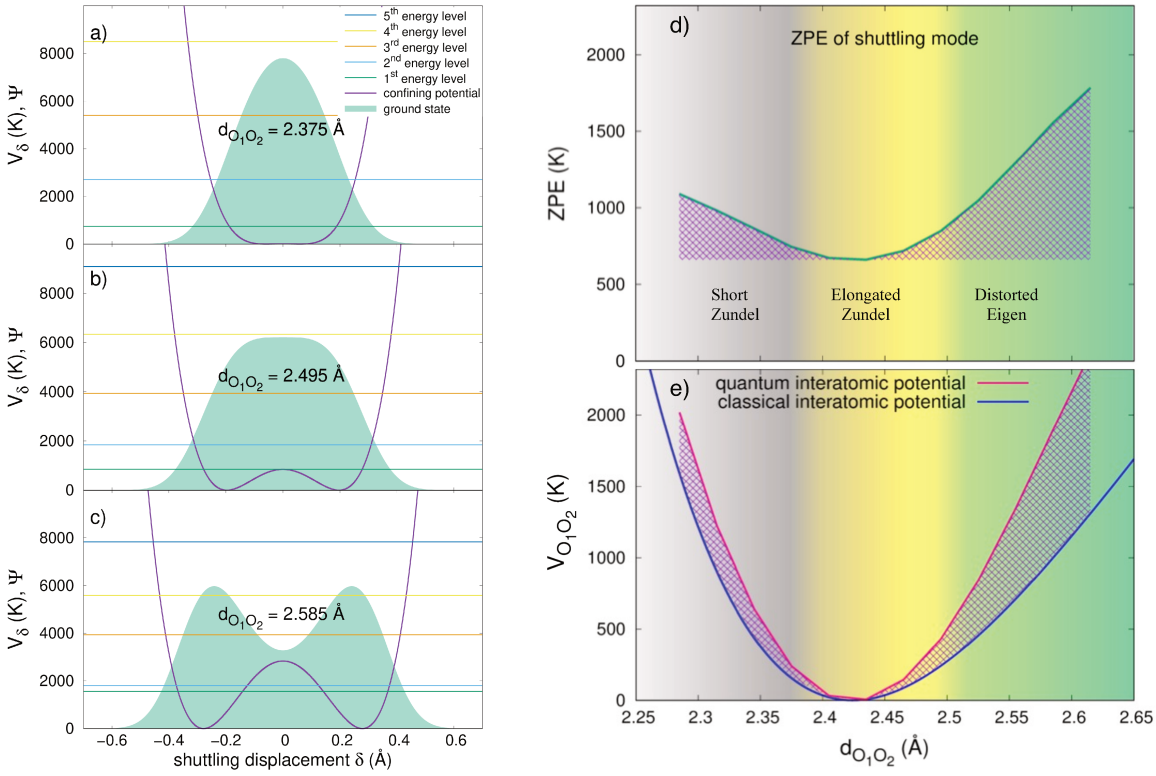


Figure 4.7: **NQEs on the shuttling mode, and their impact on the interatomic potential $V_{O_1O_2}$.** We quantize the proton shuttling mode δ_{H^+}, defined as the displacement along the segment connecting the two oxygen atoms in the core of the cluster from its mid-point position. We study the ground-state wave function and the first 5 eigenvalues for the confining potential $V_{\delta_{H^+}}$, as a function of $d_{O_1O_2}$. Panels a), b) and c) report the ground state wave function and the lowest 5 energy levels for $d_{O_1O_2}$ = 2.375, 2.495 and 2.585 Å, respectively. In panel d), the variation of the zero-point (ground-state) energy (ZPE) as a function of $d_{O_1O_2}$ is explicitly plotted. While the ZPE dependence is very flat in the elongated-Zundel region (depicted by the yellow shaded area), the ZPE increases in both short-Zundel (gray shaded area) and distorted-Eigen (green shaded area) regions, with a much steeper slope in the latter. In panel e), the ZPE is added to the classical interatomic potential $V_{O_1O_2}$ (solid blue line) to yield the quantum-corrected effective interatomic potential (solid dark-pink line) between the two inner oxygen atoms.

the ZPE decrease in the short Zundel, because it is now set by the much deeper lateral minima of the double-well potential. This can be seen again in Fig. 4.7(d).

We can now correct the classical O_1-O_2 potential, defined as $V_{O_1O_2} \equiv V_{2D}(d_{O_1O_2}, \delta_{H^+})|_{\delta_{H^+}=\delta^{min}_{H^+}}$, where $\delta^{min}_{H^+}$ is the V_{2D} minimum at fixed $d_{O_1O_2}$ value, by adding the ZPE $\forall d_{O_1O_2}$, obtained from the quantisation of the shuttling mode δ_{H^+}. The resulting potential is plotted in Fig. 4.7(e). Remarkably, the anharmonic classical $V_{O_1O_2}$ potential becomes harmonic after ZPE-correction. It is a consequence of the much larger ZPE in the distorted-Eigen configurations than in the short Zundel, which compensates for the underlying $V_{O_1O_2}$ anharmonicity. This rationalises two main features. On the one hand, it explains the very low thermal expansion of $\langle d_{O_1O_2} \rangle$, being the average position in a harmonic potential temperature-independent. On the other hand, it proves that NQEs enhance the occurrence of short-Zundel configurations upon heating, while the distorted Eigen is penalised by its large

ZPE with respect to the classical counterpart. The enhancement of the occurrence of Zundel configurations by NQEs is also revealed by the population analysis showed in the left panel of Fig. 4.8. One can see that the short-Zundel population has a peak in the [250-350] K range, in accordance with the instanton analysis of Section 4.5. Notice however that the population here is taken all over the sample, and not only over the instanton instances. Raising the temperature above the sweet spot region promotes a larger distorted-Eigen population. This is detrimental for the short-Zundel population, which indeed falls down. The maximum in the short-Zundel population corresponds to the sweet spot in the PT, showing once again the key role played by the short-Zundel species in optimizing the PT. It is interesting to study the impact of NQEs on the species population at 300 K. This is reported in the right panel of Fig. 4.8. At this temperature quantum effects favour the occurrence of the short-Zundel species with respect to the distorted Eigen states, penalised by a larger zero point energy, absent in classical calculations where the relative occurrence between the two species is reversed. This behaviour is in agreement with the outcome of Ref. [185], where a similar difference between classical and quantum populations has been found.

Above RT, the distorted Eigen configurations will eventually become dominant again. This can be understood within this framework as well. Indeed, thermal excitations are energetically more available in the distorted Eigen, where the spacing between the ZPE and the first-excited state shrinks, and higher excited states are piled up more densely than in the short and elongated Zundel (see Fig. 4.7(a-c)).

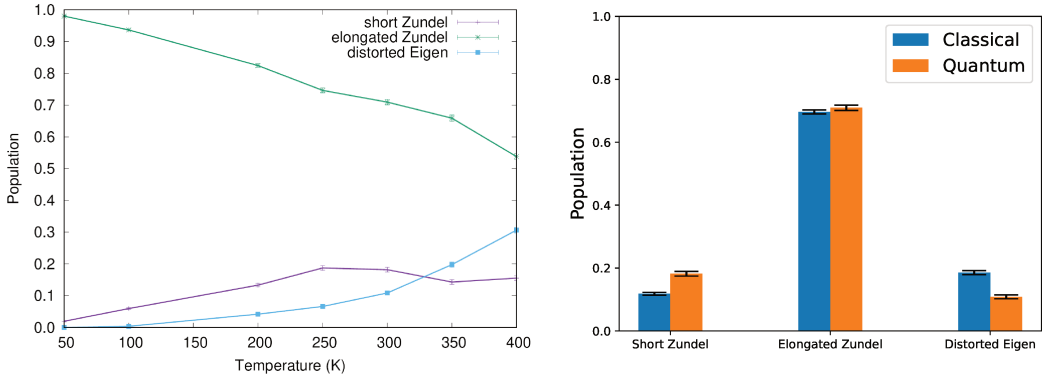


Figure 4.8: **Population of the short Zundel, elongated Zundel and distorted Eigen species**. On the left, as evaluated from the QMC-driven PIMD trajectories, and plotted as a function of temperature. The species are defined based on their $d_{O_1O_2}$ distance computed for the centroids. On the right, at 300 K, evaluated from both classical and quantum QMC-driven MD.

4.5 Optimal proton transfer from instantons statistics

The analysis made so far highlights the paramount importance of the NQEs to set the non-trivial temperature behaviour of the $H_{13}O_6^+$ cluster. At this stage, direct information about the excess

proton dynamics along the QMC-PIMD trajectory is necessary to estimate more quantitatively its impact on the PT processes occurring in the system.

One way to achieve this goal is by analysing the statistics of selected transition-state (TS) configurations, defined by means of instanton theory. Within the PI formalism, the instanton path seamlessly connects the reactants and products minima, along the minimal action trajectory, periodic in the quantum imaginary time $\tau = \beta\hbar$ [186]. It provides a generalisation of the TS theory for anharmonic quantum systems [187], and it has been very recently applied in a QMC framework [188, 189], by efficiently recovering the proper scaling of ground-state tunneling rates. TS configurations are therefore identified as those where each half of the instanton path is located on either side of the central O_1O_2 midpoint, sampled during the QMC-PIMD.

With the aim at resolving the contribution of the three different regimes to the PT dynamics, we collect the instanton events and compute their statistical distribution as a function of $d_{O_1O_2}$. We plot the instanton density distribution function in Fig. 4.9(a) at various temperatures. To deepen our analysis, we compute also the cumulative density distribution function in Fig. 4.9(b), after normalising it based on the algorithmic frequency of the instanton occurrences, as counted during our QMC-PIMD simulations. Although this does not give direct access to real-time quantities, the RPMD with Langevin thermostat has been shown to yield physically reliable information on frequencies and frequency variations [190]. Note that the coupling with the Langevin thermostat is kept constant across the full temperature range analysed here [190]. The fully integrated frequency distribution gives the total proton hopping frequency, plotted in Fig. 4.9(c) as a function of temperature. This shows a clear maximum located in the [250-300] K temperature range. Consequently, we expect the hydrated proton mobility to be optimal in a near-RT window, with a maximised Grotthus diffusion. To understand the source of this temperature "sweet spot", in the same panel (c) we plot the contribution to the total frequency of instanton events occurring in the short-Zundel region. This is yielded by the cumulative frequency distribution of panel (b) evaluated at the boundaries between short and elongated Zundel, i.e. at $d_{O_1O_2} = d_{\text{symm}}$. The short-Zundel contribution to the total frequency shows a peak of the same intensity as the total one in the same temperature range, clearly pointing to the key role played by thermally activated short-Zundel configurations to the PT dynamics. The short-Zundel arrangement enables instantaneous proton jumps between the two sides of the cation, since there is no barrier to cross. Thus, the "sweet spot" constitutes the best compromise between acquiring enough thermal energy to access short-distance configurations, boosted by NQEs, and controlling the amplitude of the chemical (covalent or H-) bonds fluctuations, that might trap the proton into an asymmetric well. Indeed, at larger temperatures (> 300 K), the onset of distorted-Eigen and the corresponding fall of short-Zundel configurations localize the hydrated proton around its closest oxygen atom, thus reducing its shuttling probability. A similar non-monotonous PT behaviour has experimentally been found in bulk water by assessing the limiting conductivities of the H_3O^+ and D_3O^+ species [191]. Thanks to these measures, performed at 20 MPa, the excess molar conductivities due to PT have been estimated. They show a peak located at a temperature in between 420 K and 430 K. In this temperature range and at the

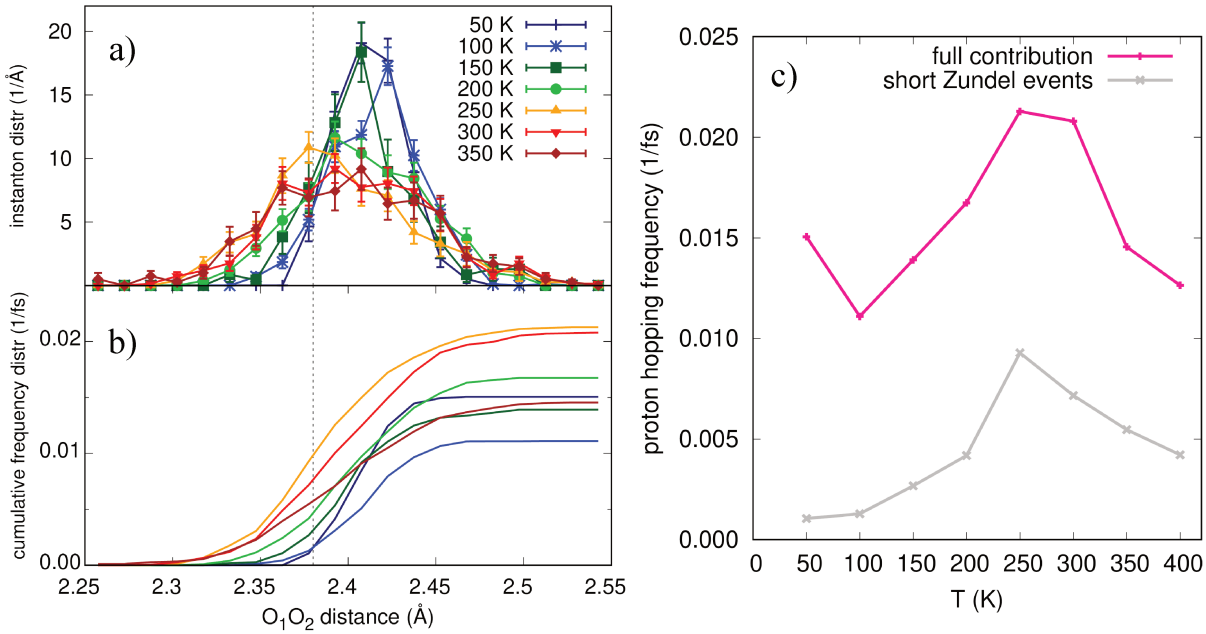


Figure 4.9: **Instanton statistics and proton hopping frequency.** a) Instanton distribution resolved as a function of the $d_{O_1O_2}$ distance for different temperatures. b) Cumulative distribution of a) normalised by the occurrence frequency of the instanton (proton hopping) events during the PIMD simulations. c) Proton hopping frequency as a function of temperature, together with the contribution coming from the short Zundel configurations, with $d_{O_1O_2} < d_{\text{symm}} = 2.38$ Å. The d_{symm} value is reported as vertical dashed line in panels a) and b). Here, we report simulations performed also at 400 K, a temperature at which the cluster is still stable or meta-stable.

pressure conditions of the experiment, the water density is only 7-8% smaller than the standard conditions [192], a regime comparable to the one of our cluster.

Beside this PT mechanism, which is adiabatic in nature and driven by the synergy of ZPE and thermal effects, NQEs could also contribute to the proton diffusion by means of instantaneous tunneling, which can further accelerate the PT dynamics. By computing the root-mean-square (RMS) displacement correlation functions[193] over the instantons population, we verified that tunneling events could take place only in the distorted Eigen and in the intermediate temperature range . This additional PT channel has however a marginal effect with respect to the main mechanism unveiled here. Indeed, Fig. 4.9(c) shows that the "sweet spot" is mainly due to PT events originating in short-Zundel configurations, where quantum tunneling is not relevant.

4.6 Discussion

Using highly accurate QMC-PIMD simulations of the $H_{13}O_6^+$ cation at finite temperature, we found a remarkably low thermal expansion of the protonated water hexamer core. It stems from a cooperative action of both NQEs and thermal effects, which leads to the emergent behaviour of short-Zundel species as PT booster, where the excess proton is perfectly shared between two neighbouring water molecules. The relevance of short-Zundel configurations is enhanced by

NQEs, which instead penalize the distorted-Eigen states, having a larger ZPE. In the intermediate temperature range, comprising RT, the occurrence of short-Zundel events is maximised by thermal population, leading to a "sweet spot" in the PT dynamics. Around these temperatures, distorted-Eigen states can still contribute to PT with quantum tunneling processes, although occurring at much lower rates. The cluster core spreads out again at larger temperatures, as soon as stronger thermal fluctuations favor the formation of more classical distorted-Eigen structures, where the proton gets strongly localised in one of the flanking molecules.

The short-Zundel quantum species is crucial for an efficient proton diffusion, as the shortness of its structure enables a fast charge redistribution during the adiabatic PT process. Recent progress in ultrafast broadband two-dimensional (2D) IR spectroscopy [194, 195] allowed to probe the vibrational properties of protonated water at vibrational frequencies around the hydrated proton stretching mode, by measuring the lowest-lying excitations in the mid-infrared continuum [195]. These state-of-the-art experiments revealed a strongly inhomogeneous behaviour of the pump-probe spectra, implying large structural distributions in proton asymmetry and O_1O_2 distance. Therefore, the traditional "Zundel limit" [18] needs to be revisited and extended, in order to cover the broad range of structures detected experimentally [196, 197]. In particular, the occurrence of qualitatively different short H-bond configurations, straightforwardly connected with the short-Zundel species described here, has been detected and highlighted in a recent fully solvated $(HF_2)^-(H_2O)_6$ experiment through femtosecond 2D IR spectroscopy in Ref. [198]. The present work crucially extends those findings by providing a temperature resolved analysis of the short H-bond events and by revealing their fundamental relation with the PT dynamics.

While proton transfer and proton transport occur in a variety of environments, from solutions to membrane proteins and fuel-cell membranes, the protonated water hexamer is one of the smallest clusters to incorporate most of the PT experimental features and solvation effects at the leading order. According to Ref. [199], one more hydration layer is needed to reach the water bulk limit. From this viewpoint, the hexamer is close to that limit, and some relevant effects, emerging already at this size, can be transferred to larger systems. Our findings thus call for further efforts to explore the temperature behaviour of the proton dynamics and transport both in aqueous systems and in other extended environments, by keeping the same accuracy as the one delivered by our QMC-driven PIMD approach in the protonated water hexamer.

In Tab. 4.1, we report the complete list of VMC+PILD simulations done for the protonated water hexamer. Owing to their importance, particularly long simulations are performed for the quantum case at temperatures of 50, 100, 200, and 300 K. In all simulations, we generated at least 1850000 electronic Monte Carlo configurations to optimise the wave function at each step of MD or PIMD. The resulting total CPU time per time step is reported in the Table. these calculations have been run on parallel machines, with two levels of parallelisation. The first one is based on the parallel sampling of the electronic degress of freedom, the second one is built upon the coupled dynamics of each beads. Notice that in our framework PIMD is not more costly than classical MD, thanks to the "bead grouping approximation" (Chapter 3).

	quantum simulations			classical simulations	
T(K)	N_{beads}	$N_{\text{iterations}}$	$t_{\text{iteration}}$(h)	$N_{\text{iterations}}$	$t_{\text{iteration}}$(h)
50	128	35282	119.4[1]	-	-
100	128	52184	24.4[2]	21454	42.0[2]
150	64	11218	-	-	-
200	64	32553	95.7[1]	20478	103.6[1]
250	32	23912	92.2[1]	24154	123.5[1]
300	32	31929	106.3[1]	22656	109.9[1]
350	32	18489	102.4[1]	26481	130.5[1]
400	32	23026	120.9[1]	27517	134.0[1]

Table 4.1: **Summary of the computational cost of the simulations on $H^+(H_2O)_6$.** In both classical and quantum calculations, a time step δt of 1 ft is used for all temperatures. The CPU time per time step ($t_{\text{iteration}}$) is also reported in hours.
[1]: calculations done on 68-core Intel Xeon Phi 7250 CPU (Knights Landing) nodes at 1.40 GHz.
[2]: calculations done on dual-processor (2x64 cores) AMD Rome (Epyc) compute nodes at 2.6 GHz.

At the moment, simulating larger structures or longer dynamics can be achieved only by overcoming the high computational cost of QMC. This goal could be achieved by training efficient atomistic machine learning potentials on QMC, which is the topic of the second Part of this thesis.

Part II

Machine learning interatomic potentials applied to quantum Monte Carlo

Chapter 5

Analytic potentials: strength and limits

Accurate simulations from first principles are quite demanding in terms of computational cost and became feasible only with contemporary advancements in hardware and software. Even today, these simulations are limited by system size and simulation time, depending on the method's scaling with the number of electrons and the intrinsic complexity of the electronic wavefunction. Consequently, before the advent of AIMD and continuing to the present, significant effort has been dedicated to developing force fields, also known as analytic potentials to stress on their functional form.

Force fields (FF) aim to provide the PES through a parametrized function that describes both intramolecular and intermolecular forces. The latter includes all interactions that do not lead to the formation of chemical bonds [200]. The absence of explicit electrons does not eliminate quantum mechanical considerations, as non-bonding interactions must still be carefully modeled to account not only for electrostatics like polarization, but also for quantum effects, such as exchange repulsion at short distances and dispersion forces over long ranges.

Depending on the origin of the data set used to fit the FF parameters, force fields can be categorized as empirical force fields (EFF), semi-empirical force fields (SEF), or force fields entirely fitted to *ab initio* data. Interestingly, even before the rise of machine learning, there was a trend towards fitting FFs more with synthetic data rather than experimental data [201]. This shift was driven not just by data availability and production costs reduced by "in silico experiments", but also by the need of force fields general enough to reproduce the quantum properties of molecules and build bottom-up explanation of chemical phenomena, rather than just casting experimental knowledge into predefined functional forms. Indeed a major drawback of EFFs that they are usually tailored to specific laboratory conditions or applications, and the parameters are often calibrated to reproduce a few specific properties, which significantly limits their prediction power and the range for exploratory work. For example if the FF has to be used in a biological context, it will be parameterized for ambient temperature and atmospheric pressure.

In this Chapter we are going to present the analytical potentials developed specifically for water in Section 5.1, and we will focus particularly on those based on many-body expansion in Section 5.2, which will be used later in Chapter 7 to produce a dataset of protonated water

clusters configurations with their respective energy and forces.

5.1 Force fields for water

Due to its fundamental role in chemistry and biology, water has become one of the most extensively modeled and parameterized compounds in computational chemistry. An exhaustive presentation of water models is beyond the scope of this thesis; for more detailed information, the reader is referred to appropriate literature [201–204]. However, some of these models are worth mentioning because they will be partially used in the following work and have paved the way towards applying functional form-agnostic and data-driven methods to the realm of interatomic potentials. Moreover, they introduced key ideas that are still used in case of new physics-aware machine learning potentials [205].

Pioneering water models based on empirical data are still used today in macroscopic simulations due to their relatively low computational cost. These classical force fields use point charges to account for Coulomb interactions, usually neglect polarization, while dispersion and repulsion typically are represented by a Lennard-Jones (LJ) term [206].

Examples include the Transferable Intermolecular Potentials based on $n = 3, 4, 5$ Point sites (TIPnP) for both charges and LJ terms [207–210], and Single Point Charges (SPC), with just three atomic sites [211]. Despite improvements with the inclusion of flexibility in water monomers (TIP4PF [212], SPC/Fw [213]), long-range electrostatics by Ewald summations (TIP4P-Ew [214]), and inclusion of quantum effects (q-TIP4P/F [215], q-SPC/Fw [216]), these empirical force fields are limited in reproducing the effects of the strong anisotropy of electronic distribution and do not account for the non-additivity of interactions.

It is only with the availability of *ab initio* data that the derivation of polarizable force fields (PFFs) became possible; PFF replace simple point charges by higher-order multipoles in order to approximate the electronic cloud, in combination with perturbation theory to rigorously describe polarization and induction [217]. Examples of models with such features, sometimes including molecular flexibility, are the Anisotropic Site Potentials (ASP) [218], the Symmetry Adapted Perturbation Theory (SAPT) water models [219], and the Thole-Type Models (TTM) [220–223]. Except for TTM, the first two typically consider only pairwise interactions, incorporating adjustments for many-body effects if necessary.

All the models mentioned above are mainly used in large simulations of water, especially in biochemical systems where macromolecules interact with a large water matrix. However, they lack two fundamental features. First, their simple analytic form does not capture the complexity of short-range quantum effects such as exchange, nor it correctly describe the electrostatic interactions and charge transfer. For example, the point-charge approximation is an oversimplification to the well known diffuse spherical charge density, and cannot correctly account for the interaction energy of overlapping charge distributions. Also the Lennard-Jones interaction term, originally proposed for closed-shell systems, is not suitable for water, leading to a large value of the first peak of the oxygen-oxygen pair correlation function [202]. Secondly, only a

few can be generalized to include reactions and interactions with water ions, and this requires significant effort. This has only been accomplished with TTM-type models [224], which explicitly account for n-body contributions. The need to combine this feature with a potential with a flexible functional form brings us to the last family of water models, the many-body expansion-based ones.

5.2 Many body expansion-based potentials for neutral and protonated water

The Many-body expansion (MBE) [225] is a fragmentation method that allows one to decompose the total energy of a molecular system as the sum of n-body contributions, where the smallest unit can be either a single atom or molecule. MBE reads as

$$E = \sum_{i=1}^{M} E_i + \sum_{i<j}^{M} \Delta E_{ij} + \sum_{i<j<k}^{M} \Delta E_{ijk} + \sum_{i<j<k<l}^{M} \Delta E_{ijkl} + \cdots \tag{5.1}$$

where M is the number of monomers. For a M-body cluster the formula expanded up to the M-body order is exact. The n-body contributions are computed as corrections to (n-1)-body ones:

$$\Delta E_{ij} = E_{ij} - E_i - E_j, \tag{5.2}$$

$$\Delta E_{ijk} = E_{ijk} - \Delta E_{ij} - \Delta E_{jk} - \Delta E_{ki} - E_i - E_j - E_k. \tag{5.3}$$

Since the number of terms scales factorially with M, in practice the expansion is usually truncated at the 3- or 4-body term, still allowing to go beyond the pairwise additivity of usual analytic potentials.

In the case of water, the smaller unit of the expansion is the single H_2O monomer, and for each *n*-body term separate PES are fitted to large datasets containing both cluster and condensed phase data computed with accurate quantum chemistry methods (MP2 and/or CC). To correctly bridge the gap in the short-range interactions between analytic potentials and *ab initio* method, highly flexible functions are employed.

Water models based on MBE are the CCpol [226–228], the dielectric polarizable point (DPP) [229], the Huang-Braams-Bowman (HBB) [230–232], and the MB-pol [233–238]. The purely-data driven potential used to fit the monomers in HBB relies on Permutationally-invariant polynomials (PIPs) [239–242]; seemingly, MB-pol uses PIPss, and has been tested with other machine learning tools [243]. MBE-based potentials gave among the best results in reproducing both cluster spectroscopic properties and bulk phase diagrams, often in combination with ML techniques, either to fit n-body terms [244] or to accelerate MBE itself [245–247]. Other contributions from n-body terms require special techniques, for which the reader is referred to a dedicated review [248].

As an example of such potentials, we focus on an improved version of HBB, WHBB [249], because it will be used later. The WHBB expansion consists of the following contributions up

to the third order:

$$V_{\mathrm{WHBB}}(1,\cdots,N) = \sum_{i=1}^{N} V_{1b}(i) + \sum_{i>j}^{N} \left[V_{2b}^{\mathrm{CCSD(T)}}(i,j) S_{2b} + V_{2b}^{\mathrm{TTM3\text{-}F}}(i,j)(1 - S_{2b}) \right] + \sum_{i>j>k}^{N} V_{3b}(i,j,k) S_{3b}, \tag{5.4}$$

where the monomer potential V_{1b} is from Partridge-Schwenke [250], the 2-body term is a function that switches between a PIPs fit to CCSD(T) energies $V_{2b}^{\mathrm{CCSD(T)}}(i,j)$ and TTM3-F interaction $V_{2b}^{\mathrm{TTM3\text{-}F}}(i,j)$ through S_{2b} depending if the two monomers i and j are in the short- or long-range regime, and the 3-body term V_{3b} is a PIPs fit to MP2 energies. Further improvements to WHBB have been proposed, namely q-AQUA [251] and q-AQUA-pol [252], but for our purposes, WHBB suffices.

The potential for water ions in the MBE framework is naively obtained by including the charged species in the expansion [253–257]. In the case of protonated water, the simplest charged monomer is the hydronium H_3O^+ ion. For example, the protonated water clusters mentioned in the first part of the thesis, namely the Zundel and of the protonated water hexamer, are represented by the following expansion:

$$V_{\mathrm{H_5O_2^+}} = V_h^{(1)} + V_w^{(1)} + V_{h,w}^{(2)}, \tag{5.5}$$

$$V_{\mathrm{H^+(H_2O)_6}} = V_{\mathrm{H_5O_2^+}} + \sum_{i=2} V_{w_i}^{(1)} + \sum_{i,j} V_{w_i,w_j}^{(2)} + \sum_{i,j} V_{h,w_i,w_j}^{(3)} + \sum_{i,j,k} V_{w_i,w_j,w_k}^{(3)} + \sum_{i,j,k} V_{h,w_i,w_j,w_k}^{(4)}, \tag{5.6}$$

proposed in [255] and [175], respectively. Here h stands for the hydronium and w_i for the i-th water molecule.

In the light of the WHBB model, the latter can also be written as:

$$V_{\mathrm{H^+(H_2O)_6}} = V_{\mathrm{H_3O^+}} + V_{\mathrm{WHBB}}((\mathrm{H_2O})_5) + \sum_{i,j} V_{h,w_i,w_j}^{(3)} + \sum_{i,j,k} V_{h,w_i,w_j,w_k}^{(4)}. \tag{5.7}$$

These protonated WHBB/q-AQUA PES have been extensively employed in the study of vibrational properties by Bowman et al. [175, 198, 256, 258–260].

Although molecular dynamics can be performed using MBE-PES, the necessity of assigning atoms to specific monomers renders the MBE approach unsuitable for modeling chemical reactions. Reactive many body expansion (RMBE), based on the sum MBE energy over all possible assignments of atoms to monomers, has been proposed specifically for the study of the protonated water hexamer [261]. Imposing a smooth distance cut-off to the interaction makes the RMBE scale polynomially with the system size.

We mention that only a few other reactive force fields have been developed, ReaXX [262] and Multistate-Empirical Valence Bond (MS-EVB) [15, 181, 183, 185, 199, 263–267], which offered many of the insights presented in the Introduction. Unfortunately MS-EVB need to be reparametrised depending on the system at hand, and it is computationally demanding.

The idea of describing atomic interaction within a fixed cutoff by means of permutationally invariant polynomials, able to fit any PES without explicit knowledge of the underlying physics but the permutational symmetries of alike atoms, anticipated key ideas of machine learning

potentials (MLIPs). Moreover, MLIPs can manage bond breaking, a necessary property to correctly describe any change in atoms assignement across different molecules, proton hopping included. MLPs will be the topic of the next Chapter.

Chapter 6

Machine learning interatomic potentials

Machine Learning (ML) is an umbrella term that refers to any partially or fully automated technique capable of identifying meaningful patterns in a given set of observations, which constitute the "experience" of the learner. Typical problems addressed by ML include classification, which involves assigning discrete labels to observed data, and regression, the continuous generalization of classification, which involves finding a mapping between dependent variables $\mathbf{y}$ and independent variables $\mathbf{X}$. ML parameters, if present, describe how the learner approaches the data provided during the training, rather than specifying a predetermined functional form expected to underlie the data. In fact the strength of these algorithms lies in their adaptability to the data, hence the term *data-driven modeling*, often used interchangeably with ML.

Fitting the Potential Energy Surface (PES) is a regression problem that can be stated as follows: given a set of $N_{\text{train}} = N$ molecular configurations containing the stoichiometry Z and the Cartesian coordinates $\mathbf{q}$ of N_{at} nuclei,

$$\{\mathbf{X}_i\}_{i=1,\cdots,N} = \{Z_i, \mathbf{q}_i\}_{i=1,\cdots N} \tag{6.1}$$

what is the functional dependence of the corresponding energy and forces

$$\{\mathbf{y}_i\}_{i=1,\cdots,N} = \{(E_i, \mathbf{f}_i)\}_{i=1,\cdots,N} \quad ? \tag{6.2}$$

This type of setting is called *supervised learning* because, during training, the algorithm has access to both the input and output data of the function it is expected to mimic.

Unfortunately, Cartesian coordinates $\mathbf{R}$ alone are not suitable for learning algorithms, because they do not transform under basic symmetry operations belonging to the group of Euclidean isometries as the energy does [268]. In fact from the molecular Hamiltonian (Eq. 1.2) it follows that the energy of a molecule is E(3)-invariant, while the Cartesian coordinates are not. E(3)-invariance means that for a transformation $\mathcal{G}$, being it a translation $\mathcal{T}$, a rotation $\mathcal{R}$ or a reflection $\mathcal{S}$, applied to the coordinates of a molecule, the energy and the forces change as follows:

$$E\left(Z, \mathcal{G}(\mathbf{R})\right) = E\left(Z, \mathbf{R}\right) \tag{6.3}$$

$$\mathbf{f}\left(Z, \mathcal{G}(\mathbf{R})\right) = \mathcal{G}\left(\mathbf{f}(Z, \mathbf{R})\right) \tag{6.4}$$

Moreover, both energy and forces are invariant under permutations $\mathcal{P}$ of identical atoms, a property that stems from the summation over atomic indices in the Hamiltonian. A cumbersome way to ensure the learning of these symmetries would be data augmentation, which involves transforming and replicating the dataset according to all the symmetries. However, this would result in an excessively large dataset.

Much of the research effort in ML for chemistry and materials science has been devoted to finding the best way to represent data in a manner that can handle symmetries. Other desirable properties of such *representations*, or descriptors, include injectivity, or *completeness*, meaning that different structures map to different descriptors, and the *differentiability*, which is necessary in order to compute the gradient of the energy and get the forces.

Descriptors can be carefully crafted based on expert knowledge—a process known as *feature engineering* in the ML community—or they can be deduced automatically, a setting referred to as *end-to-end learning*, where the algorithm itself transforms the input data.

Given the broad scope of the topic, we will review only those descriptors that will be employed in this thesis. We begin with the two main categories of representations for molecules and their PES: global representations in Section 6.1 and local representations in Section 6.2.

Then, after briefly rephrasing the regression problem in the statistical learning framework (Section 6.3), we will introduce the two big families of machine learning methods for non-linear fitting: kernel methods and neural networks in Sections 6.4 and 6.5, respectively. In each of them we will focus on the two machine learning potentials used in this work, kernel ridge regression through Operator Quantum Machine Learning (OQML) and the Message-passing neural network Atomic Cluster Expansion (MACE).

Finally, in Section 6.6 we will give an overview of what has already been done in water simulations with machine learning interatomic potentials.

6.1 Global representations

As the name suggests, global representations take full advantage of the geometry of the entire molecule. An important limitation of these descriptors is the fixed number of atoms they can consider, or the fact that, when describing compounds of multiple sizes, it is the larger one that will determine the scaling properties of the whole calculation. For the same reason, extensions for use in bulk systems is not straighforward; however, this is not a problem when fitting the PES of a single molecule.

6.1.1 Symmetrizing over pairwise distances

The Coulomb matrix [269] is worth mentioning for its simplicity as the first example of a global representation. The entries of the matrix $\mathbf{C}$ contain the reciprocal of pairwise interatomic dis-

tances multiplied by the atomic numbers Z of the pair:

$$C_{ij} = \begin{cases} \frac{1}{2}Z^{2.4}, & i = j \\ \frac{Z_I Z_J}{\|\mathbf{r}_I - \mathbf{r}_J\|}, & i \neq j \end{cases} . \tag{6.5}$$

The CM naturally incorporates the E(3)-invariances, but it is not symmetric with respect to the exchange of two atoms. This issue has been addressed for CMs specifically by sorting [270] or by using a bag-of-bonds approach [271].

The already mentioned PIPs also address this issue, but they are limited to 10 atoms [240], which makes them more suitable in combination with a many-body expansion in terms of monomers of limited size.

6.2 Local representations

We can exploit the many-body expansion introduced in the previous chapter and write the energy of a collection of N_{at} atoms as a many-atom expansion,

$$E = \sum_{a}^{N_{\text{at}}} V^{(1)}(\mathbf{R}_a) + \frac{1}{2!}\sum_{ab}^{N_{\text{at}}} V^{(2)}(\mathbf{q}_a, \mathbf{q}_b) + \frac{1}{3!}\sum_{abc}^{N_{\text{at}}} V^{(3)}(\mathbf{q}_a, \mathbf{q}_b, \mathbf{q}_c) + \cdots, \tag{6.6}$$

where the potentials terms are symmetric in the atomic positions $\mathbf{q}_a$, and zero if two or more indices are identical [272]. From the equation we can extract the single atomic contribution as

$$\mathcal{E}_a = V^{(1)}(\mathbf{q}_a) + \frac{1}{2!}\sum_{b}^{N_{\text{at}}} V^{(2)}(\mathbf{q}_a, \mathbf{q}_b) + \frac{1}{3!}\sum_{bc}^{N_{\text{at}}} V^{(3)}(\mathbf{q}_a, \mathbf{q}_b, \mathbf{R}_c) + \cdots. \tag{6.7}$$

At this point, two approximations become necessary. First, the expansion must be truncated at the Kth order. Secondly, interactions are considered only between atoms within a fixed radial cutoff, as depicted in Fig. 6.1. This significant approximation is partially justified by the principle of the nearsightedness of electronic matter (NEM), which asserts that local electronic properties are significantly influenced by the effective external potentials—such as those generated by other atoms—only at nearby points [273][1]. In both cases, the number of elements in the sum of the (K+1)-body term scales as N_c^K, where N_c is the average number of neighbors within the cutoff, making the evaluation of atomic energy computationally intensive.

This is where Machine learning interatomic potential (MLIP) come into play. Two fundamental concepts have been crucial to the successful application of data-driven approaches to interatomic potentials:

1. Interpolating the PES through local atomic contributions learned using highly flexible models.

[1]Electronic matter may be nearsighted, but it cannot be fooled! An electronic property, such as the density n, is primarily determined by the effective potential in the vicinity of that point. However, the effective potential itself, generated by nearby atoms, can be influenced by other atoms that are far away due to long-range electrostatic interactions. This brings us to the longstanding problem of including long-range interactions in MLIP [274].

2. Using n-point correlations of the atomic density instead of n-atom correlations.

The idea of point 1 is to described the local atomic environment is described using 2- and 3-body correlation-based representations, which are then used as input for highly general non-linear functions. These functions can fit virtually any type of interaction, effectively reproducing higher-order n-body terms as well. However, the impressive accuracy of these methods often comes at the cost of low interpretability. This idea was pioneered in 2007 when Behler and Parrinello proposed incorporating the permutational symmetry of atoms by considering only the atomic contributions $\mathcal{E}_a$ to the total energy E [275]:

$$E = \sum_{a}^{N_{\text{at}}} \mathcal{E}(\mathbf{x}_a) = \sum_{x}^{N_{\text{elem}}} \sum_{a}^{N_x} \mathcal{E}_x(\mathbf{x}_a), \tag{6.8}$$

where we used $\mathcal{E}_x(\mathbf{x}_a)$ to emphasize that each $\mathcal{E}_a = \mathcal{E}(\mathbf{x}_a)$ must be evaluated using the same predictive method for atoms of the same element x, considering the local atomic environment, represented by a local descriptor $\mathbf{x}_a$; we also remind that M is the total number of atoms, N_{elem} is the total number of atomic species, and N_x is the number of atoms belonging to the same element. It is important to clarify that despite the notation, x is a label for the element type, while $\mathbf{x}$ is a vector representing an atom, and these should not be confused. Once the permutational symmetry is ensured through the partition of energy among atoms, the local atomic environment descriptors must respect the symmetries of the E(3) group. In the specific case of [275], Atom Centered Symmetry Functions (ACSFs) were used as descriptors, and high-dimensional neural networks served as fitting method. Since then, local descriptors have been the subject of extensive research. The same ACSFs have been deeply explored [276], followed by many other representations, such as the Smooth Overlap of Atomic Positions (SOAP) [277], and the Faber Christensen Huang Lilienfeld (FCHL19) [278, 279].

Often these expansions of the atomic environment are truncated at the three- or four-body term, to meet efficiency needs by limiting the scaling of the descriptor evaluation to N_c^2 or N_c^3, respectively. It has been demonstrated that this truncation makes the atomic descriptor *incomplete*, meaning that injectivity requirement is not satisfied [280]. Although this does not constitute a major problem in most practical settings, it has motivated the research of a more systematic methods of including arbirary body orders of correlation. Descriptors able to do this are the moment tensor potentials (MTP) [281], the PIPs [240], and the Atomic Cluster Expansion (ACE) [282]. The latter in particular brings us to the second key idea of MLIPs, the 'density trick' [283], already anticipated by MTP and partially with SOAP. This technique is based on the fact that the n-point correlations of the atomic density around a central atom can provide a *linear* basis to expand any local property, atomic energies included [284], greatly simplyfing computation like the one in Eq. 6.7.

For complete reviews of descriptors, we refer the reader to the relevant literature [285–287]. It is worth noting that abstract approaches to descriptor design highlighted that atom-centered representations are variations of the same mathematical object [288], and that they can be re-

duced to truncated expansion of the more general ACE framework. Moreover, this similarities have been made apparent also in message-passing frameworks [284].

Since in this thesis we are going to employ several descriptors (FCHL19, ACSFs and those automatically built through message-passing extending from ACE), we will write down the main ideas behind ACE [282] and FCHL19 [279].

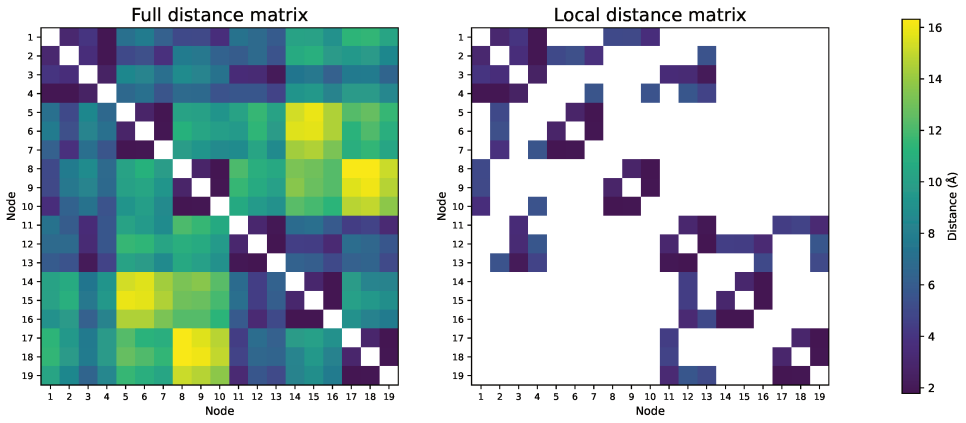


Figure 6.1: **Comparison between full distance matrix and local atomic environments in the protonated water hexamer.** On the left the distance matrix of the protonated water hexamer, on the right its masked representation when only atomic environment with $r_{\text{cut}} = 3\,\text{Å}$ are considered. Atoms are numbered as in Fig. 6.9.

6.2.1 Atomic cluster expansion

Consider an atom a surrounded by N_c atoms falling into a sphere with given radius r_{cut}. Its atomic environment is fully described by

$$\sigma_{\mathbf{q}} = \left\{\mathbf{r}_{1a}, z_1, \mathbf{r}_{2a}, z_2, \cdots, \mathbf{r}_{N_c a}, z_{N_c}\right\}, \tag{6.9}$$

where each relation with a neighbouring atom b is fully characterized by the vector separating a and b, $\mathbf{r}_{ba} = \mathbf{q}_b - \mathbf{q}_a$, and atomic species of b, z_b, as in Fig. 6.2. For the sake of readability, we will drop the a that refers to the central atom.

The atomic cluster expansion allow to systematically decompose a property of an atom a, like its energy

$$\mathcal{E}_a(\sigma) = \mathcal{E}_a(\mathbf{r}_{1a}, \mathbf{r}_{2a}, \cdots, \mathbf{r}_{N_c a}), \tag{6.10}$$

in contributions coming from each element of the powerset of its neighbouring atoms. By grouping different subsets of σ having the same number of elements K, we get (K+1)-body contributions.

ACE is based on progressive definition of basis functions: first the single-bond basis, then cluster one and finally the atomic basis. The single-bond basis is made of functions that fully describe the central atom and a neighbouring one. Considering the variables $\mathbf{r} = \mathbf{r}_{ba}$ and $z_b = z$ for short, an element of this basis is given by:

$$\phi_{xv}(\mathbf{r}, z) := \delta(z - x)\varphi_{nlm}(\mathbf{r}) \tag{6.11}$$

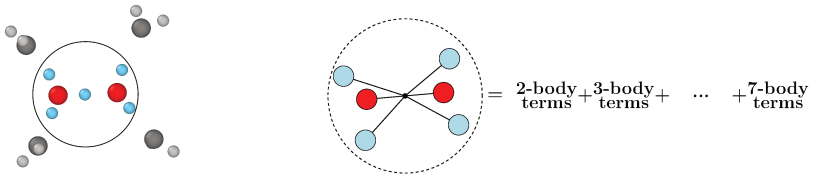


Figure 6.2: **Local energy as sum of n-body terms.** We aim to describe the local atomic contribution to the total energy. Here we consider the central proton in the hexamer as an example. For simplicity we choose a radial cutoff such that it contains only the adjacent water molecules, for a total of $N_c = 6$ atoms in the surrounding.

where the index x serves to distinguish among different chemical elements basis, and v is a collective index in place of those of a spatial function φ_{nlm}. The latter is defined as a product between a radial part R_{nl} and a spherical harmonic Y_{lm} describing the angular part:

$$\varphi_{nlm}(\mathbf{r}) = R_n(r)Y_{lm}(\hat{\mathbf{r}}), \tag{6.12}$$

From this functions one can build a complete and possibly orthogonal basis:

$$\sum_z \int d\mathbf{r}\phi^*_{\alpha v}(\mathbf{r},z)\phi_{\beta u}(\mathbf{r},z) = \delta_{uv}\delta_{\alpha\beta} \tag{6.13}$$

$$\sum_{vx} \phi^*_{vx}(\mathbf{r},z)\phi_{vx}(\mathbf{r}',z') = \delta(\mathbf{r}-\mathbf{r}')\delta_{zz'}. \tag{6.14}$$

An atomic environment can be described as a collection of clusters μ of K atoms $\{b_j\}_{j=1,\cdots,K}$, excluded the central one, a. Consider clusters of fixed size K. Then the bonds will be $\mu = (b_1a, b_2a, \cdots, b_Ka)$ and their respective single-bond indices can be collected in the list $\nu = (v_1, v_2, \cdots, v_K)$. The cluster basis around the atom a can be built from products of single-bond functions and can be indexed with μ and ν:

$$\Phi_{\mu\nu}(\boldsymbol{\sigma}) = \varphi_{v_1}(\mathbf{r}_{b_1a})\varphi_{v_2}(\mathbf{r}_{b_2a})\cdots\varphi_{v_K}(\mathbf{r}_{b_Ka}) = \prod_{i=1}^{K}\varphi_{v_i}(\mathbf{r}_{b_ia}), \tag{6.15}$$

If there are more chemical species N_{elem}, the bonds list should not change under permutation alike atoms and can be rewritten simply as $\mu = (k_{x_1}, k_{x_2}, \cdots, k_{x_{N_{\text{el}}}})$, where $k_{x_1} + k_{x_2} + \cdots + k_{x_{N_{\text{el}}}} = K$, while the cluster basis is

$$\Phi_{\mu\nu}(\boldsymbol{\sigma}) = \Phi_{k_xk_y\cdots,\nu}(\boldsymbol{\sigma}) = \prod_{x=1}^{N_{\text{el}}}\prod_{i=1}^{k_x}\phi_{xv_i}(\mathbf{r}_{b_ia}, z_{b_i}). \tag{6.16}$$

The cluster basis will inherit the completeness and orthogonality from the single-bond basis, allowing to expand a local property like the local atomic energy as

$$\mathcal{E}_a(\boldsymbol{\sigma}) = \sum_{\mu\nu} J_{\mu\nu}\Phi_{\mu\nu}(\boldsymbol{\sigma}) \tag{6.17}$$

with the expansion coefficients obtained by projection $J_{\mu\nu} = \left\langle \Phi_{\mu\nu} | E_a(\sigma) \right\rangle$. The sum run over all possible cluster sizes k, and it is possible to rewrite it in a single-bond representation to make it more similar to the original Eq. 6.7:

$$\begin{aligned}
\mathcal{E}_a(\sigma) = &\sum_b \sum_x \sum_v J^{(1)}_{xv} \phi_{xv}(\mathbf{r}_{ba}) \\
&+ \frac{1}{2!} \sum_{b_1<b_2} \sum_{x_1 x_2} \sum_{v_1 v_2} J^{(2)}_{x_1 v_1 x_2 v_2} \phi_{x_1 v_1}(\mathbf{r}_{b_1 a}) \phi_{x_2 v_2}(\mathbf{r}_{b_2 a}) \\
&+ \frac{1}{3!} \sum_{b_1<b_2<b_3} \sum_{x_1 x_2 x_3} \sum_{v_1 v_2 v_3} J^{(3)}_{x_1 v_1 x_2 v_2 x_3 v_3} \phi_{x_1 v_1}(\mathbf{r}_{b_1 a}) \phi_{x_2 v_2}(\mathbf{r}_{b_2 a}) \phi_{x_3 v_3}(\mathbf{r}_{b_3 a}) \\
&+ \cdots .
\end{aligned} \tag{6.18}$$

To make the step to the atomic basis easier, it is possible to rewrite the above expansion with unrestricted sums with new coefficients, c. These are defined in such a way that self-interaction terms involving products of more single-bonds basis function on the same atom, for example $\phi_{x_1 v_1}(\mathbf{r}_{b_1 a}) \phi_{x_2 v_2}(\mathbf{r}_{b_1 a})$, are zero.

$$\begin{aligned}
\mathcal{E}_a(\sigma) = &\sum_b \sum_x \sum_v c^{(1)}_{xv} \phi_{xv}(\mathbf{r}_{ba}) \\
&+ \frac{1}{2!} \sum_{b_1 b_2} \sum_{x_1 x_2} \sum_{v_1 v_2} c^{(2)}_{x_1 v_1 x_2 v_2} \phi_{x_1 v_1}(\mathbf{r}_{b_1 a}) \phi_{x_2 v_2}(\mathbf{r}_{b_2 a}) \\
&+ \frac{1}{3!} \sum_{b_1 b_2 b_3} \sum_{x_1 x_2 x_3} \sum_{v_1 v_2 v_3} c^{(3)}_{x_1 v_1 x_2 v_2 x_3 v_3} \phi_{x_1 v_1}(\mathbf{r}_{b_1 a}) \phi_{x_2 v_2}(\mathbf{r}_{b_2 a}) \phi_{x_3 v_3}(\mathbf{r}_{b_3 a}) \\
&+ \cdots .
\end{aligned} \tag{6.19}$$

Until now we just wrote the original many-atom expansion as linear combination of single-bonds terms, but we did not solve the problem of the N_c^K scaling. The trick for this is to just reorder the summations such that the sum over all neighbors b is done first. This is equivalent to define an *atomic basis* as the projection of the density of atoms of element z in the neighborhood of the central atom onto single-bond basis functions:

$$A_{a,vx} = \langle \rho^x_a | \phi_{vx} | = \rangle \sum_{b:z_b=x} \phi_{vx}(\mathbf{r}_{ba}) \tag{6.20}$$

where the density of atoms of element x is defined as

$$\rho^x_a(\mathbf{r}) = \sum_b \delta_{z_b x} \delta(\mathbf{r} - \mathbf{r}_{ba}), \tag{6.21}$$

Then the atomic energy becomes a polynomial in A

$$\begin{aligned}
\mathcal{E}_a(\sigma) = &\sum_x \sum_v c^{(1)}_{xv} A_{a,vx} \\
&+ \sum_{x_1 x_2} \sum_{v_1 \le v_2} c^{(2)}_{x_1 v_1 x_2 v_2} A_{a,v_1 x_1} A_{a,v_2 x_2} \\
&+ \frac{1}{3!} \sum_{x_1 x_2 x_3} \sum_{v_1 \le v_2 \le v_3} c^{(3)}_{x_1 v_1 x_2 v_2 x_3 v_3} A_{a,v_1 x_1} A_{a,v_2 x_2} A_{a,v_3 x_3} \\
&+ \cdots .
\end{aligned} \tag{6.22}$$

A visual example of this procedure is represented in Fig. 6.3.

A further step in the ACE expansion is imposing the symmetries we mentioned at the beginning of the Chapter. For example rotational invariance may be imposed by restricting the sum over the spherical harmonics indices $\nu = (v_1 v_2)$ such that only products of spherical harmonics that can be reduced to a representation of the identity of the rotation group. This is done by reducing the products of the atomic bases $A_{a,nlm}$ using Clebsch-Gordan coefficients (or the analogous Wigner 3j symbols), which imposes conditions on the values of l m.

$$B^{(1)}_{an} = A_{an00}, \tag{6.23}$$

$$B^{(2)}_{an_1n_2l} = \sum_{m=-l}^{l} (-1)^m A_{an_1lm} A_{an_2l-m}, \tag{6.24}$$

$$B^{(3)}_{\substack{an_1n_2n_3 \\ l_1l_2l_3}} = \sum_{m_1=l_1}^{l_1} \sum_{m_2=l_2}^{l_2} \sum_{m_3=l_3}^{l_3} \begin{bmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \end{bmatrix} A_{an_1l_1m_1} A_{an_2l_2m_2} A_{an_3l_3m_3}, \tag{6.25}$$

The final energy can be written as:

$$\mathcal{E}_a(\sigma) = \sum_n c^{(1)}_n B^{(1)}_{an} + \sum_{n_1n_2l} c^{(2)}_{n_1n_2l} B^{(2)}_{an_1n_2l} + \sum_{\substack{n_1n_2n_3 \\ l_1l_2l_3}} c^{(3)}_{\substack{n_1n_2n_3 \\ l_1l_2l_3}} B^{(3)}_{\substack{an_1n_2n_3 \\ l_1l_2l_3}} + \cdots \tag{6.26}$$

or, in a more compact form:

$$E_a(\sigma) = \sum_{Knl} \mathbf{c}^{(K)}_{\mathbf{nl}} \mathbf{B}^{(K)}_{\mathbf{anl}} \tag{6.27}$$

6.2.2 Faber Christensen Huang Lilienfeld (FCHL19) descriptor

As for all the atom-centered representations, the molecule m is represented by the collection $\mathbf{X}$ of the descriptors $\mathbf{x}$ of all the N_{at} atoms belonging to it.

$$\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_{N_{\text{at}}}] \tag{6.28}$$

If we denote a generic atom with a, its environment is described by two types of symmetry functions (other terms can be added, but they would decrease the performances):

$$\mathbf{x}_a = [\mathbf{G}^{\textbf{2-body}}_{\mathbf{a}}, \mathbf{G}^{\textbf{3-body}}_{\mathbf{a}}], \tag{6.29}$$

where we dropped the single molecule superscript m. Each of the two term is a collection on its own:

- Two-body radial functions, $\mathbf{G}^{\textbf{2-body}}_{\mathbf{a}}$, describe the distribution of chemical elements around the central atom a. It is physically related to the coordination number and it scales linearly with the number of elements N_{elem} of possible elements in the atomic environment. In

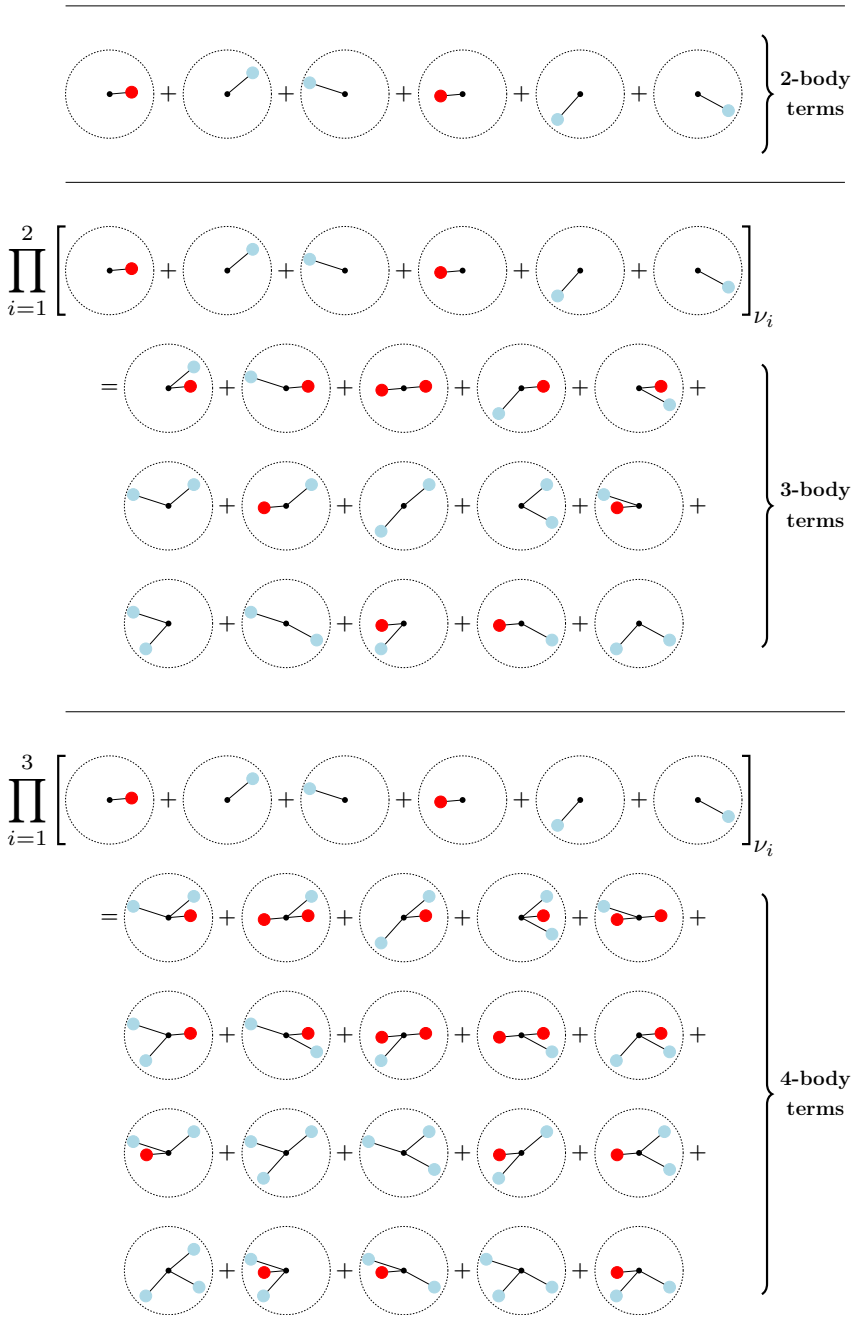


Figure 6.3: **Graphical explanation of ACE density trick.** In general the energy will be the sum of n-body terms contribution, up to the $(N_c + 1)$th order, which includes all the atoms in the local environment. If we consider combination of single-bonds function to form the n-body terms, allowing repetitions, each (K+1)-body term will contain N_c^K contributions, which can be costly to evaluate, considering that this has to be done for each atom in the molecule. A more convenient way of evaluating the energy is by using products of the atomic basis (2nd and 6th lines), which scales linearly with the number of neighbours. Once the atomic basis is defined, it can be employed for higher-order contributions to the total energy (2-, 3- and higher body contribution), which do not depends on the number of nearby atoms.

fact it is defined as a collection of element-wise radial functions $G_a^{\text{2-body}}(x_\alpha)$ where x_α is one of the possible elements:

$$\mathbf{G_a^{2\text{-body}}} = [\underbrace{G_a^{\text{2-b}}(x_1), \cdots, G_a^{\text{2-b}}(x_{N_{\text{elem}}})}_{N_{\text{elem}}}], \tag{6.30}$$

where each term is given by the sum of all the contributions coming from all the atoms of the same chemical species, and it is based on the reciprocal distance only:

$$G_a^{\text{2-body}}(x) = \sum_{b:Z_b=Z_x} G^{\text{2-body}}(r_{ab}) \tag{6.31}$$

where r_{ab} is the reciprocal distance between atoms a and b.

- Three-body functions, $\mathbf{G_a^{3\text{-body}}}$, describes the distribution of angles and distances between triplet of elements around the central atom (with the species of the central kept fixed). The scaling with the number of possible elements is given by all the possible non-ordered combination with repetition of chemical elements:

$$\left(N_{\text{elem}} + \frac{N_{\text{elem}}(N_{\text{elem}}-1)}{2!}\right) = \frac{1}{2}N_{\text{elem}}(N_{\text{elem}}+1) \tag{6.32}$$

$$\mathbf{G_a^{3\text{-body}}} = [\underbrace{G_a^{\text{3-b}}(x_1,x_1), G_a^{\text{3-b}}(x_1,x_2), \cdots, G_a^{\text{3-b}}(x_1,x_{N_{\text{el}}}), G_a^{\text{3-b}}(x_2,x_2), \cdots, G_a^{\text{3-b}}(x_{N_{\text{el}}},x_{N_{\text{el}}})}_{N_{\text{el}}(N_{\text{el}}+1)}], \tag{6.33}$$

where each term is a shorthand for

$$G_a^{\text{3-b}}(x,x') = \sum_{\substack{b:Z_b=Z_x \\ c:Z_c=Z_{x'}}} G^{\text{3-b}}(R_s, r_{ab}, r_{ac}, \theta_{abc}, \theta_{cab}, \theta_{bca}), \tag{6.34}$$

as before we have the sum over different contribution coming from couple of atoms in the atomic environment, taking into account the reciprocal distance with the central atom and the three angles formed by the triplet.

Now we have a closer look to the specific functional form of the 2- and 3-body function introduced so far in the summation.

Two-body functions

The radial basis functions set is defined over a grid, centered on the considered atom, of n_{Rs2} points, that is, the discrete variable R_s can assume n_{Rs2} values (24 by default), from $\frac{r_{\text{cut}}}{n_{Rs2}}$ to a cutoff radius r_{cut} above which the environment is not considered anymore local. The height of the bin at each R_s is given by the sum of all the contributions of the kind $G^{\text{2-body}}(r_{ab})$ coming from all the atoms belonging to the same chemical element:

$$G^{\text{2-body}}(r_{ab}) = \xi_2(r_{ab}) f_{\text{cut}}(r_{ab}) \frac{1}{R_s \sigma(r_{ab})} e^{-\frac{(\ln R_s - \mu(r_{ab}))^2}{2\sigma(r_{ab})^2}}, \tag{6.35}$$

where $\mu(r_{ab})$ and $\sigma(r_{ab})$ are parameters of the log-normal distribution; these parameters depend on the interatomic distance, r_{ab}, and a hyper-parameter, w, defined as follows:

$$\mu(r_{ab}) = \ln\left(\frac{r_{ab}}{\sqrt{1+\frac{w}{r_{ab}^2}}}\right), \tag{6.36}$$

$$\sigma(r_{ab})^2 = \ln\left(1+\frac{w}{r_{ab}^2}\right), \tag{6.37}$$

In the equation (6.35), the form of the two body scaling function, $\xi_2(r_{ab})$ has been found by previous studies to be suitable for obtaining higher regression weights to terms that contribute the most to the total energy

$$\xi_2(r_{ab}) = \frac{1}{r_{ab}^{N_2}}. \tag{6.38}$$

The soft cut-off function used here is the same as the one proposed in other representations such as ACSFs:

$$f_{\text{cut}}(r_{ab}) = \begin{cases} \frac{1}{2}\left(\cos\left(\frac{\pi r_{ab}}{r_{\text{cut}}}\right)+1\right) & \text{if} \quad r_{ab} \leq r_{cut} \\ 0 \quad \text{if} \quad r_{ab} > r_{cut}. \end{cases} \tag{6.39}$$

All the hyper-parameters (the width parameter of the log-normal distribution, w; the exponent of the scaling function, N_2; the cut-off distance, r_{cut}; and the number of radial basis functions, n_{Rs2}) have been optimized on different datasets through Monte Carlo by the authors, but can in principle be adapted to specific datasets, at the price of loosing their universal validity. Their actual values will be reported in Appendix C.

Three-body functions

The three-body functions encode the distances of an atom to neighboring pairs of atoms in the environment of the atom, as well as the angle between the triplet. The resulting function is a product of the following terms:

$$\begin{aligned} G^{\text{3-b}}(r_{ab}, r_{ac}, \theta_{abc}, \theta_{cab}, \theta_{bca}) = \\ \xi_3 G^{\text{3-body}}_{\text{radial}}(r_{ab}, r_{ac}) G^{\text{3-body}}_{\text{angular}}(\theta_{cab}) f_{\text{cut}}(r_{ab}) f_{\text{cut}}(r_{ca}) f_{\text{cut}}(r_{bc}), \end{aligned} \tag{6.40}$$

where there is a radial basis function defined as:

$$G^{\text{3-body}}_{\text{radial}}(r_{ab}, r_{ac}) = \sqrt{\frac{\eta_3}{\pi}} \exp\left(-\eta_3\left(\frac{1}{2}(r_{ab}+r_{ac}) - R_s\right)^2\right), \tag{6.41}$$

where η_3 is a parameter that controls the width of the radial distribution functions and again R_s is the location of the radial gridpoints, in total n_{Rs3} (20 by default). The three-body scaling function, ξ_3 is

$$\xi_3 = c_3 \frac{1 + 3\cos(\theta_{cab})\cos(\theta_{abc})\cos(\theta_{bca})}{(r_{ab} r_{bc} r_{ca})^{N_3}}, \tag{6.42}$$

here any θ_{ABC} is the angle $A\hat{B}C$. Finally, the angular term $G^{\text{3-body}}_{\text{angular}}$ collects two forms:

$$G^{\text{3-body}}_{\text{angular}}(\theta_{cab}) = \begin{cases} G^{\cos}_n(\theta_{cab}) = \exp\left(-\frac{(\varsigma n)^2}{2}\right)(\cos(n\theta_{cab}) - \cos(n(\theta_{cab}+\pi)) \\ G^{\sin}_n(\theta_{cab}) = \exp\left(-\frac{(\varsigma n)^2}{2}\right)(\sin(n\theta_{cab}) - \sin(n(\theta_{cab}+\pi))), \end{cases} \tag{6.43}$$

where n_F is the order of expansion (usually fixed to 1) and ς is a hyper-parameter describing the width of the angular Gaussian function.

Length of the representation of an atomic environment

According to the number of 2- and 3-body functions, N_{elem} and $N_{\text{elem}}(N_{\text{elem}}+1)$ respectively, the number of bins in each single 2- and 3-body function (n_{Rs2} and n_{Rs3} respectively) and the order of expansion of the angular term (n_F), the total length of the atomic environment descriptor is:

$$N_{\text{elem}} \times n_{Rs2} + N_{\text{elem}} \times (N_{\text{elem}}+1) \times n_{Rs3} \times n_F \tag{6.44}$$

For example, in the case of water clusters, we have only 2 types of elements, $N_{\text{elem}} = 2$, and using the default values of the expansion numbers $(n_{Rs2}, n_{Rs3}, n_F) = (24, 20, 1)$, the description of each atomic environment is a vector of 168 entries, showed as an example in Fig. 6.4

Relation to ACE

As showed in [282], it is possible to find connection between the ACSFs and the ACE descriptors. Since FCHL19 is a variation on Behler's ACSFs, we can apply the same logic here. First of all the 2-body functions of FCHL19 are just radial function that can be used in any other descriptor, included the $B^{(1)}$ term in ACE. In the 3-body functions the angular dependence is given by $\cos(\theta_{cab})$; an analogous dependence on the cosine of the angle between 2 atoms and the central one can be easily obtained from $B^{(2)}$ by means of the addition theorem for spherical harmonics:

$$\frac{4\pi}{2l+1}\sum_{m=-l}^{l}(-1)^m Y_l^m(\hat{\mathbf{r}}_{ba}) Y_l^{-m}(\hat{\mathbf{r}}_{ca}) = P_l(\cos(\theta_{cab})) \tag{6.45}$$

where P_l are Legendre polynomials.

6.3 Regression in the statistical learning framework

For a more comprehensive and in-depth exploration of the statistical learning framework, we refer to [289, 290].

We start from a set of observations formed by couples of independent and dependent variables

$$\underbrace{(\mathbf{x}_1, \mathbf{y}_1), \cdots, (\mathbf{x}_{N_{\text{train}}}, \mathbf{y}_{N_{\text{train}}})}_{\text{Train set}} \in \mathcal{X} \times \mathcal{Y}, \tag{6.46}$$

where $\mathcal{X}$ is the domain set, or sample/instance space, and $\mathcal{Y}$ is called label or target set; in our case they would be the molecule representations and the corresponding tuples of energy and forces $(E, \mathbf{F})$, respectively.

In the statistical learning framework the regression problem is rephrased as: assuming that the data are generated according to an unknown joint distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, which we can write as $\mathcal{D} = \mathcal{D}((\mathbf{x}, \mathbf{y})|\mathbf{x})\mathcal{D}(\mathbf{x})$, what is the function (called also predictor, or model) f that minimizes the probability $\mathbb{P}$ of sampling $\mathbf{x}$ and missing its right target $\mathbf{y}$, i.e.

$$\mathbb{P}[f] := \mathcal{D}(\{(\mathbf{x}, \mathbf{y}) : f(\mathbf{x}) \neq \mathbf{y}\})? \tag{6.47}$$

where $\mathbb{P}$ is a functional whose domain is all the functions belonging to an hypothesis class $\mathcal{H}$. This can be recast into the minimization of the true risk, $L_{\mathcal{D}}$, which is defined through this same probability.

The true risk can not be evaluated, since we do not know $\mathcal{D}$ and we do not have a disposal an infinite number of couples $(\mathbf{x}, \mathbf{y})$ to compute it. A more practical way to measure the success

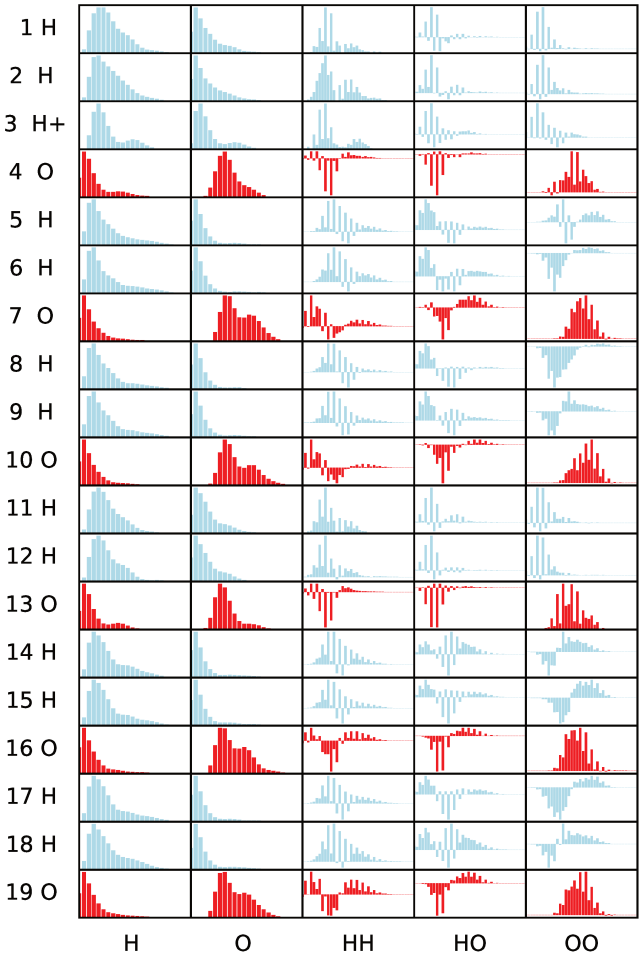


Figure 6.4: **FCHL19 descriptor for a $H^+(H_2O)_6$ configuration.** Each row represent an single atomic environment of 168 entries, distributed in columns of different element-wise n-body terms: the first two are 2-body terms (24 bins each), the last three are the 3-body terms (20 bins each), containing angular information. We can see how the third hydrogen, which is the proton, is the only one showing a double peak 2-body correlation with both oxygens and hydrogens. Overall difference can be noticed also between the central Zundel oxygens, at the 4th and 13th columns, with respect to all the solvation oxygens, and between hydrogens in the solvation shell and Zundel hydrogens.

of a predictor is through the empirical risk minimization (ERM) learning rule:

$$\min_f \left[L_s \left[(f(\mathbf{x}_i), \mathbf{y}_i)_{i=1,\cdots,n} \right] \right] = \min_f \left[\frac{1}{N} \sum_{i=1}^{N} l_s \left(f(\mathbf{x}_i), \mathbf{y}_i \right) \right], \tag{6.48}$$

where l_s is an appropriate loss function which measure the error given a true value $\mathbf{y}_i$ and the prediction $f(\mathbf{x}_i)$, while L_s is the total loss. However in most of the cases the regularized loss minimization (RLM) learning rule is adopted, where the functional to be minimized is:

$$\min_f \left[L_s(f) + R(f) \right], \tag{6.49}$$

where the first term is the empirical risk defined above, $R(f)$ is the regularization term which controls the complexity of the hypothesis class $\mathcal{F}$ from which f is selected: it should be large enough to contain the functions that can solve the problem, but not too large, otherwise the algorithm could overfit the data or be unstable under slight change of its input.

An example is the regularized least-squares linear regression, where the unknown function $f_{\mathbf{w}}$ is approximated by the hyperplane that minimizes the squared distance between the predicted and the true function value.

$$\min_{\mathbf{w}} \left[\frac{1}{N} \sum_{i=1}^{N} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2 + \lambda \left\| \mathbf{w} \right\|^2 \right], \tag{6.50}$$

where λ is a hyper-parameter which controls the trade-off between high empirical risk or high complexity. Now the predictor function $f_{\mathbf{w}}(\mathbf{x})$ is defined as a dot product between the instance $\mathbf{x}$ and the vector of coefficients $\mathbf{w}$, as usual in linear regression; the loss function is the squared loss, the regularization function is the Tikhonov one. This kind of regression is called ridge regression.

6.4 Kernel methods

Here, we introduce Kernel ridge regression (KRR) methods. We refer to [290, 291] for a broader and deeper view on these subjects.

6.4.1 Kernel ridge regression

Kernel methods extend what we have seen above with linear regression to nonlinear functions. The samples $\{\mathbf{x}_i\}$ are mapped to a high-dimensional Hilbert space called *feature space*[2], $\mathcal{H}$, where the learning task can be reduced to a linear regression. The explicit feature map can be formally defined as

$$\begin{aligned} \phi: \quad & \mathcal{X} \to \mathcal{H} \\ & \mathbf{x} \mapsto \mathbf{q} := \phi(\mathbf{x}), \end{aligned} \tag{6.51}$$

[2]NB: these features are different from those produced in the features engineering step we introduced in Sec. 6.1 and 6.2; that one in our case means "descriptor engineering"

given the high number of dimensions of feature space (up to infinite), it can be costly if not impossible to compute the features of a given sample $\mathbf{x}$. The "kernel trick" is a way around this problem: if we rewrite the problem of (6.50) in a more general way by considering the vectors belonging to the feature space we obtain

$$\arg\min_{\mathbf{w}} \left[L_s \left(\langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle, \cdots \langle \mathbf{w}, \phi(\mathbf{x}_N) \rangle; \mathbf{y} \right) + R(\|\mathbf{w}\|) \right], \tag{6.52}$$

we can apply the representer theorem [292], which states that there exists a vector $\boldsymbol{\alpha} \in \mathbb{R}^n$ such that the optimal solution of the equation (6.52) can be written as:

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i \phi(\mathbf{x}_i). \tag{6.53}$$

This allows one to rewrite the optimization problem (6.52) as:

$$\arg\min_{\boldsymbol{\alpha}} \left[L_s \left(\sum_{j=1}^{N} \alpha_j \langle \phi(\mathbf{x}_j), \phi(\mathbf{x}_1) \rangle, \cdots, \sum_{j=1}^{N} \alpha_j \langle \phi(\mathbf{x}_j), \phi(\mathbf{x}_N) \rangle; \mathbf{y} \right) + R \left(\sqrt{\sum_{ij}^{N} \alpha_i \alpha_j \langle \phi(\mathbf{q}_j), \phi(\mathbf{q}_i) \rangle} \right) \right], \tag{6.54}$$

in which we notice that the features appear only in the dot product in the feature space. In the particular case of regularized least squares we have

$$\operatorname{argmin}_{\boldsymbol{\alpha}} \left[\sum_{i}^{N} \left(\sum_{j=1}^{N} \alpha_j \langle \phi(\mathbf{x}_j), \phi(\mathbf{x}_i) \rangle - y_i \right)^2 + \lambda \sum_{ij}^{N} \alpha_i \alpha_j \langle \phi(\mathbf{x}_j), \phi(\mathbf{q}_i) \rangle \right]. \tag{6.55}$$

Now we define the kernel function $\mathcal{K}$ as

$$\begin{aligned} \mathcal{K} : & \mathcal{X} \times \mathcal{X} \to \mathbb{R} \\ & (\mathbf{x}, \mathbf{x}') \mapsto \mathcal{K}(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}'), \phi(\mathbf{x}) \rangle, \end{aligned} \tag{6.56}$$

which implies that the kernel must be symmetric, that is $\mathcal{K}(\mathbf{x}, \mathbf{x}') = \mathcal{K}(\mathbf{x}', \mathbf{x})$, and positive definite. Equation (6.54) can be written as an optimization problem with respect to the coefficients $\boldsymbol{\alpha}$,

$$\operatorname{argmin}_{\boldsymbol{\alpha}} \left[L_s \left(\sum_{j=1}^{N} \alpha_j \mathcal{K}(\mathbf{x}_j, \mathbf{x}_1), \cdots, \sum_{j=1}^{N} \alpha_j \mathcal{K}(\mathbf{x}_j, \mathbf{x}_N); \mathbf{y} \right) + R \left(\sqrt{\sum_{ij}^{N} \alpha_i \alpha_j \mathcal{K}(\mathbf{x}_j, \mathbf{x}_i)} \right) \right], \tag{6.57}$$

for which we do not need direct access to the elements in the features space $\mathcal{H}$ through the explicit mapping ϕ, we need only to know how to perform the inner product $\mathcal{K}(\mathbf{x}, \mathbf{x}')$, that is, the kernel function, or equivalently, the Gram matrix $\mathbf{K}$:

$$K_{ij} = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle. \tag{6.58}$$

Particularly for the least-squares, the Gram matrix $\mathbf{K}$ allows one to write (6.55) in a compact way by constructing the label vector $\mathbf{y}^T = (y_1, \cdots, y_N)$ and the vector of coefficients $\boldsymbol{\alpha}^T = (\alpha_1, \cdots, \alpha_N)$:

$$\operatorname{argmin}_{\boldsymbol{\alpha}} \left[\sum_{i}^{N} \left(\sum_{j=1}^{N} \alpha_j K_{ij} - y_i \right)^2 + \lambda \sum_{ij}^{N} \alpha_i \alpha_j K_{ij} \right], \tag{6.59}$$

$$\underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \left[\|\mathbf{K}\boldsymbol{\alpha} - \mathbf{y}\|^2 + \lambda \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} \right]. \tag{6.60}$$

Once we learn the coefficients $\boldsymbol{\alpha}$, we can calculate the prediction on a new instance $\mathbf{x}^*$ by simply computing the dot product:

$$\langle \mathbf{w}, \phi(\mathbf{x}^*) \rangle = \sum_j^N \alpha_j \langle \phi(\mathbf{x}_j), \phi(\mathbf{x}^*) \rangle = \sum_j^N \alpha_j \mathcal{K}(\mathbf{x}_j, \mathbf{x}^*). \tag{6.61}$$

Before treating the specific case of energy learning, we mention one of the most commonly used kernel, which will be used also later, the Gaussian one, also called radial basis function (RBF):

$$\begin{aligned}
\mathcal{K}(\mathbf{x}, \mathbf{x}') &= e^{-\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{2\sigma^2}} = e^{-\frac{(\mathbf{x}-\mathbf{x}')^2}{2\sigma^2}} = e^{-\frac{\mathbf{x}^2}{2\sigma^2} - \frac{\mathbf{x}'^2}{2\sigma^2} + \frac{\mathbf{x}\mathbf{x}'}{\sigma^2}} = \\
&= e^{-\frac{\mathbf{x}^2}{2\sigma^2} - \frac{\mathbf{x}'^2}{2\sigma^2}} \left(1 + \frac{\mathbf{x}\mathbf{x}'}{\sigma^2} + \frac{1}{2!} \left(\frac{\mathbf{x}\mathbf{x}'}{\sigma^2} \right)^2 + \frac{1}{3!} \left(\frac{\mathbf{x}\mathbf{x}'}{\sigma^2} \right)^3 + \cdots \right) = \\
&= e^{-\frac{\mathbf{x}^2}{2\sigma^2} - \frac{\mathbf{x}'^2}{2\sigma^2}} \left(1 \cdot 1 + \frac{\mathbf{x}}{\sigma} \cdot \frac{\mathbf{x}'}{\sigma} + \frac{1}{\sqrt{2!}} \frac{\mathbf{x}^2}{\sigma^2} \cdot \frac{1}{\sqrt{2!}} \frac{\mathbf{x}'^2}{\sigma^2} + \frac{1}{\sqrt{3!}} \frac{\mathbf{x}^3}{\sigma^3} \cdot \frac{1}{\sqrt{3!}} \frac{\mathbf{x}'^3}{\sigma^3} + \cdots \right) = \\
&= \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle,
\end{aligned} \tag{6.62}$$

where the passages makes evident that we have implicitly employed the mapping:

$$\phi(\mathbf{x}) = e^{-\frac{\mathbf{x}^2}{2\sigma}} \left[1, \frac{\mathbf{x}}{\sigma}, \frac{1}{\sqrt{2!}} \frac{\mathbf{x}^2}{\sigma^2}, \frac{1}{\sqrt{3!}} \frac{\mathbf{x}^3}{\sigma^3}, \cdots \right]. \tag{6.63}$$

6.4.2 Learning energies via kernel methods

Now we can apply the concepts developed in the previous section to the specific problem of fitting the PES. In the case of global representations, we can write the training set as

$$(\mathbf{X}_1, E_1), \cdots, (\mathbf{X}_n, E_n), \cdots, (\mathbf{X}_N, E_N) \tag{6.64}$$

The training step consists in solving the minimization problem (6.55) in its matrix inversion problem form (Eq. 6.60), where $\mathbf{K}$ is a square matrix of shape $N_{\text{train}} \times N_{\text{train}}$.

Instead of just inverting the equation according to a naive loss minimization, it is common practice in ML to reduce the complexity of the hypothesis class of the function, in this case represented by the regression coefficients, by introducing a penalty for too large coefficients. This learning rule is called regularized loss minimization (RLM). This is done also to stabilize the learning algorithm, which means that a slight change of its input should not change too much its output. One of the most employed regularization function is the Tikhonov one we have already seen in Eq. (6.50):

$$R(\boldsymbol{\alpha}) = \lambda \|\boldsymbol{\alpha}\|_2^2 = \lambda \sum_n^N \alpha_n^2, \tag{6.65}$$

In our case the kernel ridge regression is done by solving the minimization problem:

$$\underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \left[\frac{1}{2} \|\mathbf{K}\boldsymbol{\alpha} - \mathbf{E}\|_2^2 + \frac{\lambda}{2} \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} \right] \tag{6.66}$$

The solution can be written in the following closed-form:

$$\boldsymbol{\alpha} = \left(\mathbf{K} + \mathbb{I}\boldsymbol{\lambda}\right)^{-1} \mathbf{E}. \tag{6.67}$$

In practice this solution is usually determined not by direct inversion of the matrix, but by singular value decomposition, to deal with the large dimension of the training set.

Once the coefficients are found, the total energy of a query molecule m would be given by

$$E_m^* = E^*(\mathbf{X}_m) = \sum_n^N K_{mn}\alpha_n = \sum_n^N \mathcal{K}(\mathbf{X}_m^*, \mathbf{X}_n)\alpha_n. \tag{6.68}$$

where the α's are the regression coefficients, and the kernel matrix entries are given by a kernel based function, which for example could be

$$K_{mn}^* = \mathcal{K}(\mathbf{X}_n, \mathbf{X}_m^*) = \exp\left(-\frac{\|\mathbf{X}_n - \mathbf{X}_m^*\|_2^2}{2\sigma} \right) \tag{6.69}$$

If we want to predict the energies of $N_{\text{test}} = M$ molecules, we can collect them in a single vector $\mathbf{E}^* = [E_1^*, E_2^*, \cdots, E_M^*]^T$ and express the above equation in matrix form:

$$\mathbf{E}^* = \mathbf{K}^* \boldsymbol{\alpha} \tag{6.70}$$

If we are describing the compound by local atomic environments $\mathbf{x}$, these are compared by locally defined kernels; the total energy of a molecule is now expressed as a sum of local atomic energies:

$$E_m^* = \sum_{a \in m} \mathcal{E}(x_a^*) = \sum_{a \in m} \sum_{n=1}^{N} \sum_{b \in n} \mathcal{K}(\mathbf{x}_b^*, \mathbf{x}_a)\alpha_n \tag{6.71}$$

where m and n are the test and train configuration indices, and for short notation we indicated with $a \in m$ and $b \in n$ the atoms belonging to them. If we rearrange the sum we can still consider a "global kernel", given by the sum of local ones:

$$E_m^* = \sum_{n=1}^{N} \sum_{a \in m} \sum_{b \in n} \mathcal{K}(\mathbf{x}_b, \mathbf{x}_a^*)\alpha_n = \sum_{n=1}^{N} K_{mn}^* \alpha_j, \tag{6.72}$$

and as we did for the true global kernel in Eq. 6.70, we can write the above equation in algebraic form, $\mathbf{E}^* = \mathbf{K}^* \boldsymbol{\alpha}$. In both cases the kernel matrix $\mathbf{K}^*$ has a shape of $M \times N$. However, in the latter case, the matrix entries K_{mn} do not correspond to true kernel functions. In fact this compact form hides the sum over the atoms used to build the basis. Using again RBF as an example, we can write

$$K_{mn}^* = \sum_{a \in m} \sum_{b \in n} \mathcal{K}(\mathbf{x}_b, \mathbf{x}_a^*) = \sum_{a \in m} \sum_{b \in n} \delta_{Z_b, Z_a} \exp\left(\frac{\|\mathbf{x}_b - \mathbf{x}_a^*\|_2^2}{2\sigma} \right). \tag{6.73}$$

where the Kronecker-delta δ_{Z_b, Z_a} has been introduced to compute the local kernel only between atoms of the same type, following the prescription of Eq. (6.8).

6.4.3 Gaussian process regression kernel

What has been exposed so far can be studied also from a Bayesian view point. For a general treatment of the subject see [293], for its application to PES learning see [294]. In particular Gaussian process regression (GPR) offers an alternative to fit simultuneously energies and forces, with kernels of the form:

$$\begin{bmatrix} \mathbf{E} \\ \mathbf{F} \end{bmatrix} = \begin{bmatrix} \mathbf{K} & -\frac{\partial}{\partial \mathbf{r}^T}\mathbf{K} \\ -\frac{\partial}{\partial \mathbf{r}}\mathbf{K} & \frac{\partial^2}{\partial \mathbf{r}\partial \mathbf{r}^T}\mathbf{K} \end{bmatrix} \boldsymbol{\alpha}. \tag{6.74}$$

where the dimensions are $(N + 3N_{\text{at}}N) \times (N + N_{\text{at}}N)$, with the usual N the number of samples in the training set and N_{at} the number of atoms in each sample. Now the cost function is:

$$\underset{\boldsymbol{\alpha}}{\text{argmin}} \left[\frac{1}{2} \left\| \begin{bmatrix} \mathbf{K} & -\frac{\partial}{\partial \mathbf{r}^T}\mathbf{K} \\ -\frac{\partial}{\partial \mathbf{r}}\mathbf{K} & \frac{\partial^2}{\partial \mathbf{r}\partial \mathbf{r}^T}\mathbf{K} \end{bmatrix} \boldsymbol{\alpha} - \begin{bmatrix} \mathbf{E} \\ \mathbf{F} \end{bmatrix} \right\|_2^2 + \frac{\lambda}{2} \boldsymbol{\alpha}^T \begin{bmatrix} \mathbf{K} & -\frac{\partial}{\partial \mathbf{r}^T}\mathbf{K} \\ -\frac{\partial}{\partial \mathbf{r}}\mathbf{K} & \frac{\partial^2}{\partial \mathbf{r}\partial \mathbf{r}^T}\mathbf{K} \end{bmatrix} \boldsymbol{\alpha} \right] \tag{6.75}$$

with solution:

$$\boldsymbol{\alpha} = \left(\begin{bmatrix} \mathbf{K} & -\frac{\partial}{\partial \mathbf{r}^T}\mathbf{K} \\ -\frac{\partial}{\partial \mathbf{r}}\mathbf{K} & \frac{\partial^2}{\partial \mathbf{r}\partial \mathbf{r}^T}\mathbf{K} \end{bmatrix} + \mathbb{I}\lambda \right)^{-1} \begin{bmatrix} \mathbf{E} \\ \mathbf{F} \end{bmatrix} \tag{6.76}$$

Energies and forces can be obtained with:

$$\mathbf{E} = \left[\mathbf{K}, -\frac{\partial}{\partial \mathbf{r}^T}\mathbf{K} \right] \boldsymbol{\alpha} \tag{6.77}$$

$$\mathbf{F} = -\frac{\partial}{\partial \mathbf{r}}\mathbf{E} = \left[-\frac{\partial}{\partial \mathbf{r}}\mathbf{K}, -\frac{\partial^2}{\partial \mathbf{r}\partial \mathbf{r}^T}\mathbf{K} \right] \boldsymbol{\alpha} \tag{6.78}$$

It is also possible to use only forces labels during the training to produce accurate and still energy-conserving molecular force fields. This process goes under the name of gradient-domain machine learning (GDML) [295], the equations are formally similar to the previous ones, just restricted to the lower-right $(N + 3N_{\text{at}}N) \times (N + 3N_{\text{at}}N)$ submatrix:

$$\mathbf{F} = \left[\frac{\partial^2}{\partial \mathbf{r}\partial \mathbf{r}^T}\mathbf{K} \right] \boldsymbol{\alpha} \tag{6.79}$$

$$\underset{\boldsymbol{\alpha}}{\text{argmin}} \left[\frac{1}{2} \left\| \frac{\partial^2}{\partial \mathbf{r}\partial \mathbf{r}^T}\mathbf{K} - \mathbf{F} \right\|_2^2 + \frac{\lambda}{2} \boldsymbol{\alpha}^T \frac{\partial^2}{\partial \mathbf{r}\partial \mathbf{r}^T}\mathbf{K}\boldsymbol{\alpha} \right] \tag{6.80}$$

$$\boldsymbol{\alpha} = \left(\left[\frac{\partial^2}{\partial \mathbf{r}\partial \mathbf{r}^T}\mathbf{K} \right] + \mathbb{I}\lambda \right)^{-1} \mathbf{F}. \tag{6.81}$$

The only difference is that now the energies are predicted up to an integration constant c, which can be useful only when doing the final test with direct comparison of energy values:

$$\mathbf{E} = \left[-\frac{\partial}{\partial \mathbf{r}^T}\mathbf{K} \right] \boldsymbol{\alpha} + c, \tag{6.82}$$

which in our case has been determined during the training by computing the average difference between the predicted energies and the true ones.

6.4.4 Local kernels with operator quantum machine learning

Within the context of kernel-based regression, it has been proposed [296] that not only the energies can be used as labels, but also any observable that is related to differential operators acting on the energy. This framework is named Operator Quantum Machine Learning (OQML) and it is useful in those cases in which forces are available, since it has been shown [297] that they can improve the prediction both in energy and forces.

The kernel matrix looks different from the previous one, as now the column index runs not just on the configurations in the training set, but on all the atoms of all the training set configurations:

$$K_{iJ}^{\mathrm{OML}} = \sum_{I \in i} \mathcal{K}(\mathbf{x}_J, \mathbf{x}_I) \tag{6.83}$$

which is no more a square-matrix, as the dimensions of $\mathbf{K}^{\mathrm{OML}}$ are $N \times MN$. To take into account also the forces we consider the derivative of the kernel entry with respect to the coordinate of the atoms belonging to the configuration-row:

$$-\frac{\partial}{\partial r_K^*} K_{iJ}^{\mathrm{OML}} = -\sum_{I \in i} \frac{\partial}{\partial r_K} \mathcal{K}(\mathbf{x}_J, \mathbf{x}_I) \qquad K \in [1, \cdots, 3M], \tag{6.84}$$

where the derivative is computed as shown in the previous section and K runs up to $3M$ because we are considering a particular configurations. We can express the least-squares algorithm in matrix form as the minimization of the cost function:

$$\underset{\boldsymbol{\alpha}}{\mathrm{argmin}} \left[\left\| \begin{bmatrix} \mathbf{E} \\ \mathbf{F} \end{bmatrix} - \begin{bmatrix} \mathbf{K}^{\mathrm{OML}} \\ -\frac{\partial}{\partial \mathbf{r}^*} \mathbf{K}^{\mathrm{OML}} \end{bmatrix} \boldsymbol{\alpha}^{\mathrm{OML}} \right\|_2^2 \right] \tag{6.85}$$

where the dimensions of the kernel matrix, derivatives part included, are $(3MN + N) \times MN$, with N the number of samples in the training set and M the number of atoms in each sample. Here there is no regularization factor, but since this equation is solved with singular value decomposition (SVD), the threshold below which singular value are no more considered can be treated as λ.

6.5 Neural networks

Another way of managing non-linearity is through neural networks, which have been inspired by how the brain works. The idea of modeling network of neurons dates back to 1943 [298], while the first physical implementation of a single artificial neuron able to learn to distinguish pictures is Rosenblatt's perceptron [299]. Both the biological and artificial neuron are depicted in Fig. 6.5. The biological neuron receives various inputs through chemical signals at the dendrites. If a certain threshold of signals is reached, the message is electrically propagated through the axon and passed to other neurons at the synapses. Similarly, the perceptron receives an input vector of values, and if a weighted sum of these values exceeds a threshold, it spikes an output signal, encoded as a Heaviside step-function or as another non-linear function. A single

neuron, however, is limited in its expressive power and can only solve problems that are linearly separable [300]. It is by connecting multiple perceptrons that we create a neural network, which for this reason is also called multilayer-perceptron.

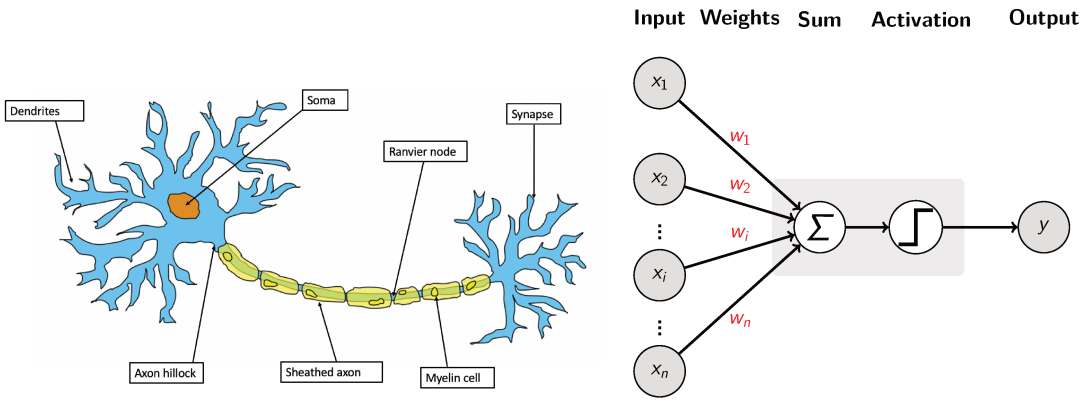


Figure 6.5: **Modeling the neuron.** On the left a biological neuron [301], on the right its artificial implementation in the perceptron

A feedforward neural network (FFNN), schemed in Fig. 6.6, is a predictor composed by:

- A directed acyclic graph $G = (V, E)$, where the nodes V are the neurons processing the information and the edges E are the link propagating it between neurons.
- A weight function $w : E \to \mathbb{R}$, which assign a weight to each edge.
- An activation function $\sigma : \mathbb{R} \to \mathbb{R}$, which model each neurons non-linear response in every nodes. Some examples are step functions $\text{sign}(x)$, or sigmoid functions $(1 + e^{-x})^{-1}$.

The network is organized in layers V_t, meaning that the set of nodes can be decomposed into a union of disjoint non-empty subsets,

$$V = \bigcup_{t=0}^{T} V_t, \tag{6.86}$$

where T is the depth of the network, that is the number of layers such that every edge in E connects some node in V_{t-1} to some node in V_t, for some $t \in [T]$. V_0 is the input layer, it contains $m + 1$ neurons, where m is the dimensionality of the input space; this means that $\forall i \in [m]$ the output of neuron i in V_0 is simply x_i. The last neuron of the first layer, $i = m + 1 = |V_0| = d_0$, is the constant neuron, which always output 1. The layers $V_1, \cdots, V_{T-1}$ are called hidden layers, and the last one, V_T, is the output layer, which gives the prediction. Each layers has a width $d_t = |V_t|$, and the total size the networks is given by the number of nodes, $|V|$, while its width is given by the largest layer, $\max_t |V_t| = \max_t d_t$.

The neural network is called feedforward because the information flows in one direction—from input to output—without any cycle or loop. At the level of a single neuron at the index i of a layer t, as sketched in Fig. 6.7, its input $a_i^{(t)}$ is the weighted sum of the outputs of all the

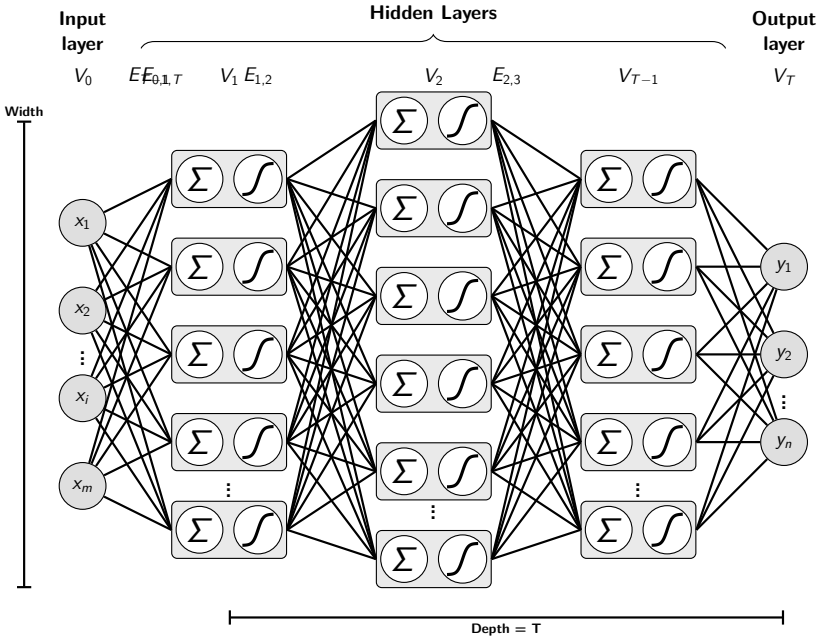


Figure 6.6: **Scheme of an artificial neural network.**

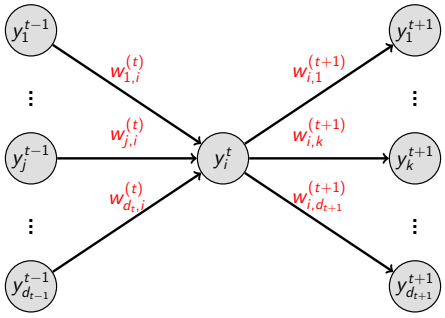


Figure 6.7: **Information flow in a single neuron embedded in a feedforward neural network.**

neurons connected to it from the layer $(t-1)$:

$$a_i^{(t-1)} = \sum_{j=1}^{d_{t-1}} \left(w_{ji}^{(t-1)}\right)^T y_j^{(t-1)} = (\mathbf{w}_i^{(t-1)})^T \mathbf{y}^{(t-1)} \tag{6.87}$$

where in the last step we expressed the sum as a matrix-vector product. The output of a neuron is simply the application of the activation function to the input:

$$y_i^{(t)} = \sigma\left((\mathbf{w}_i^{(t-1)})^T \mathbf{y}^{(t-1)} + b^{(t-1)}\right), \tag{6.88}$$

where $b^{(t-1)}$ is the bias of the neuron

We can also adopt a layer point of view, by defining the weight matrix $\mathbf{W}_t$ of shape $d_t \times d_{t-1}$, whose rows are the transposed weight vectors $\mathbf{w}_j^{(t)}$ between layer $(t-1)$ and layer (t),

$$\mathbf{W}_t = \left[\mathbf{w}_1^{(t)}, \mathbf{w}_2^{(t)}, \cdots, \mathbf{w}_{d_t}^{(t)}\right]^T. \tag{6.89}$$

It follows that we can write the total input from layer $(t-1)$ to layer t as:

$$\mathbf{a}^{(t)} = \mathbf{W}_t \mathbf{y}^{(t-1)} \tag{6.90}$$

and the total output from layer t to layer $(t+1)$ as:

$$\mathbf{y}^{(t)} = \sigma\left(\mathbf{W}_t \mathbf{y}^{(t-1)}\right) \tag{6.91}$$

Since we will deal with only FFNN, we will refer to them simply as Neural network (NN).

From a functional point of view, a NN is a function with the following domain of applications:

$$f_{V,E,\sigma,w} : \mathbb{R}^{|V_0|-1} \to \mathbb{R}^{|V_T|}. \tag{6.92}$$

where the parameters $\{V, E, \sigma\}$ define the architecture of the neural network. The hypothesis class of a network is defined by fixing its architecture

$$\mathcal{H}^{\mathrm{NN}}_{V,E,\sigma} = \{f_{V,E,\sigma,w} : w : E \to \mathbb{R}\} \tag{6.93}$$

where w is the function that assign a weight to each node. Once the hypothesis class is defined, we can denote the neural network as $f_{\mathbf{w}}$, where $\mathbf{w}$ stands for all the weights, which are the parameters to be learned. Hence training the NN means finding the optimal set of weights that minimize the loss.

6.5.1 Optimization by gradient descent

Differently from kernels, the optimization of a NN with respect to a loss function does not have a closed form solution. Therefore we rely on iterative procedures, which minimize the loss function by taking its gradient with respect to the learning parameters, here the weights:

$$\mathbf{w}^{[n+1]} = \mathbf{w}^{[n]} - \eta_n \nabla_{\mathbf{w}} L(f_{\mathbf{w}}), \tag{6.94}$$

where η_n is an adaptive learning rate, which decreases at each iteration n according to a power law or an exponential law; the idea is that at the beginning it is advisible to update the weights spanning the most possible of the loss function landscape, and avoid slow convergence; but closer to the minimum the update of the weights should be smaller, to avoid instability. Such algorithms are collectively designated as gradient descent algorithms, and will eventually land on a local minimum of the loss function L.

In order to avoid the iteration to be stuck in local minima of the loss function, stochastic gradient descent (SGD) algorithm are commonly employed [302]. The training set is randomly partitioned in minibatches, and at each iteration the gradient is computed on a single randomly picked minibatch. Other benefits of these methods may include a sort of regularization that prevents overfitting [303]. One of the most common SGD algorithm is the Adaptive Moment Estimation (ADAM) [304], which exploits first and second momenta of the gradient to calculate an adaptive learning rate.

In the case of NNs, the partial derivative with respect to a weight $w_{ji}^{(t)}$ is computed as:

$$\frac{\partial L}{\partial w_{ji}^{(t)}} = \frac{\partial L}{\partial a_i^{(t)}} \frac{\partial a_i^{(t)}}{\partial w_{ji}^{(t)}} = \Delta_i^{(t)} \frac{\partial}{\partial w_{ji}^{(t)}} \left(\sum_{j'=1}^{d_{t-1}} w_{j'i}^{(t)} y_{j'}^{(t)} \right) = \Delta_i^{(t)} y_j^{(t-1)} \tag{6.95}$$

where we see that between layer $(t-1)$ and (t), $\Delta_i^{(t)}$ can be expressed recursively as a function of the weights of the successive layer $(t+1)$:

$$\Delta_i^{(t)} = \frac{\partial L}{\partial a_i^{(t)}} = \frac{\partial L}{\partial y_i^{(t)}} \frac{\partial y_i^{(t)}}{\partial a_i^{(t)}} = \left(\sum_{k=1}^{d_{t+1}} \frac{\partial L}{\partial a_k^{(t+1)}} \frac{\partial a_k^{(t+1)}}{\partial y_i^{(t)}} \right) \sigma' \left(a_i^{(t)} \right) = \left(\sum_{k=1}^{d_{t+1}} \Delta_k^{(t+1)} w_{ik}^{(t+1)} \right) \sigma' \left(a_i^{(t)} \right) \tag{6.96}$$

with the initial condition being just the gradient of the loss function with respect to the weights of the last layer,

$$\Delta_i^T = \frac{\partial L}{\partial a_i^{(T)}}. \tag{6.97}$$

This means that in order to find the gradient of the loss function with respect to all the weight, Eq. 6.97 has to be backpropagated [305]. Once the gradient is determined, it can be used in SGD algorithms to optimize the weights.

6.5.2 Neural networks for PES

We mention that in 1990s the first NNs applied to PES fitting were global in nature [306–309]. Since the seminal paper of Behler and Parrinello [275], with a few exceptions [310], nearly all neural network potentials (NN-PES) relied on the local atomic environment approximation discussed in Sec. 6.2. The only global NN-PES nowdays are based on techniques to construct permutationally invariant global basis function, for example PIPs+NN [311]. For an historical overview of neural network potentials, we refer to [312]. In the following Section, we will focus on the two architecture employed in this thesis.

6.5.3 High-dimensional neural networks

High-dimensional neural network (HDNN) are a collection of disjointed sub-FFNNs designed to compute the atomic contributions to the energy and the forces acting on individual atoms. Only the output of these sub-networks are combined to yield the total energy of a given molecule. Despite the name, the single elemental sub-networks are relatively shallow compared to typical image processing NNs. Usually, they consists of only 2-3 layers, while the width of the networks rarely exceeds 40 nodes. The name 'high-dimensional' likely refers to the breathrough concept of using multiple sub-networks, as illustrated in Fig. 6.8. In the literature, this architecture is often referred to as Behler-Parrinello neural networks (BPNNs).

While the training process via backpropagation of the loss function's derivatives with respect to the weights 6.95 is well known, an often overlooked feature of these networks is how the forces are computed in the forward propagation step. Rather than calculating numerically

the atomic forces via finite difference, these forces can be obtained analytically using the forward derivative with respect to the descriptors. For example, the forces can be derived from the sum of local energies as:

$$\mathbf{f} = -\boldsymbol{\nabla} E = -\boldsymbol{\nabla} \sum_{a}^{N_{at}} \mathcal{E}(\mathbf{x}_a). \tag{6.98}$$

If we focus on a single atom a force component f_a^α, we can apply the chain rule [276]:

$$f_a^\alpha = -\sum_{a}^{N_{at}} \frac{\partial \mathcal{E}(\mathbf{x}_a)}{\partial q_a^\alpha} = -\sum_{a}^{N_{at}} \frac{\partial \mathcal{E}(\mathbf{x}_a)}{\partial \mathbf{x}_a} \frac{\partial \mathbf{x}_a}{\partial q_a^\alpha}. \tag{6.99}$$

The partial derivatives of the descriptors with respect to the Cartesian coordinates can be computed explicitly, or by means of symbolic or automatic differentiation. The partial derivatives of the local energies with respect to the input vector $\mathbf{x}$ of the NN instead are given by the matrix product of the Jacobians of each layer:

$$\frac{\partial \mathcal{E}}{\partial \mathbf{x}} = \mathbf{J}_T(\mathbf{y}_{T-1}) \mathbf{J}_{T-1}(\mathbf{y}_{T-2}) \cdots \mathbf{J}_1(\mathbf{x}) \tag{6.100}$$

where each Jacobian $\mathbf{J}_t$ is given by:

$$\mathbf{J}_t(\mathbf{y}_t) = \mathbf{diag}\left[\sigma'(\mathbf{W}_t \mathbf{y}_{t-1} + \mathbf{b}_t)\right] \mathbf{W}_t. \tag{6.101}$$

Alternatively, each Jacobian with respect to the input $\mathbf{x}$ can also be defined recursively as:

$$\mathbf{J}_t(\mathbf{x}) = \mathbf{diag}\left[\sigma'(\mathbf{W}_t \mathbf{y}_{t-1}(\mathbf{x}) + \mathbf{b}_t)\right] \mathbf{W}_t \mathbf{J}_{t-1}(\mathbf{x}) \tag{6.102}$$

where $\mathbf{J}_1(\mathbf{x}) = \mathbf{I} \in \mathbb{R}^{d_1 \times d_0}$ is the identity matrix, as usual in the forward derivative definition in automatic differentiation.

These computations are essential for obtaining forces and run molecular dynamics with MLIPs, and are common also in ML as input sensitivity analysis [313].

6.5.4 Graph neural networks and MACE

Structural formulas in chemistry suggest that the most natural way to represent a molecule for mathematical analysis is as a graph, $G = (V, E)$. While we already introduced graphs in the previous subsection on neural networks, in this context, each node $v \in V$ represents an atom, and the edges between two nodes $(u, v) \in E$ are undirected and purely based on the distance matrix D, rather than on actual chemical bonds. An example of such a graph for the protonated water hexamer is shown in the left panel of Fig. 6.9.

Graphs are a good starting point for machine learning chemistry, because the desired permutational symmetry translates to node-order equivariance, which is a fundamental property of graphs. Predictions on a graph can be of three types: predictions on nodes, predictions on edges and global prediction on the overall structure. Our focus is on predictions at the node level—specifically, the atoms—since we are interested in local atomic contribution to the total energy and in the forces acting on atoms. Additionally, we aim at computing the descriptor of

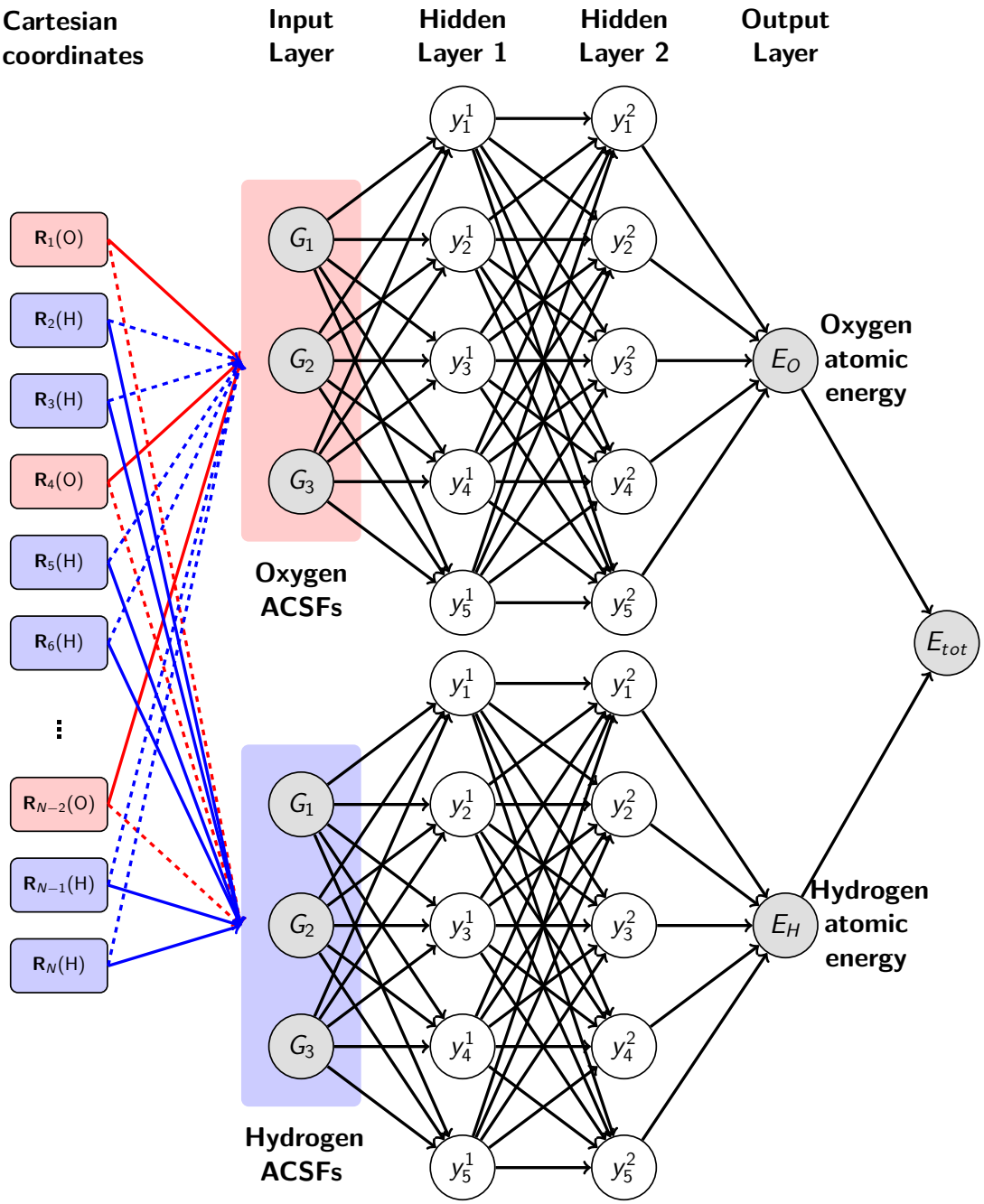


Figure 6.8: **Scheme of a HDNN for a water system.** A combination of the schemes in [314] (under the CC-BY), where we stress that there are separated network to process the environments of different atomic species. A structure is completely defined by its cartesian coordinates and its atomic species, reported in the first column. The second column is the input layer of the NN and it is formed by the atomic environment descriptors, which are a collection of Atom Centered Symmetry Functions (ACSFs), $[G_1, G_2, G_3]$; the central atom is linked to its respective environment with a full line, in fact we distinguish between oxygen descriptors, $\mathbf{x}_{\mathrm{O}}$ (in red), and hydrogen descriptors, $\mathbf{x}_{\mathrm{H}}$ (in blue); in principle all the atoms are considered when building the descriptors, hence all the cartesian coordinates are linked to the input layer (see the dashed lines of oxygens going to the descriptors of hydrogens, and vice versa); in practice, the atoms that are outside the radial cutoff will not contribute in defining $\mathbf{x}$. From the input layer we have the usual feedforward propagation through a shallow network of 2-3 hidden layers, that determines the atomic energy contribution of each atom, in the output layer. Summing over all the E^a_{O} and E^a_{H} gives us the total energy of the molecule, E_{tot}.

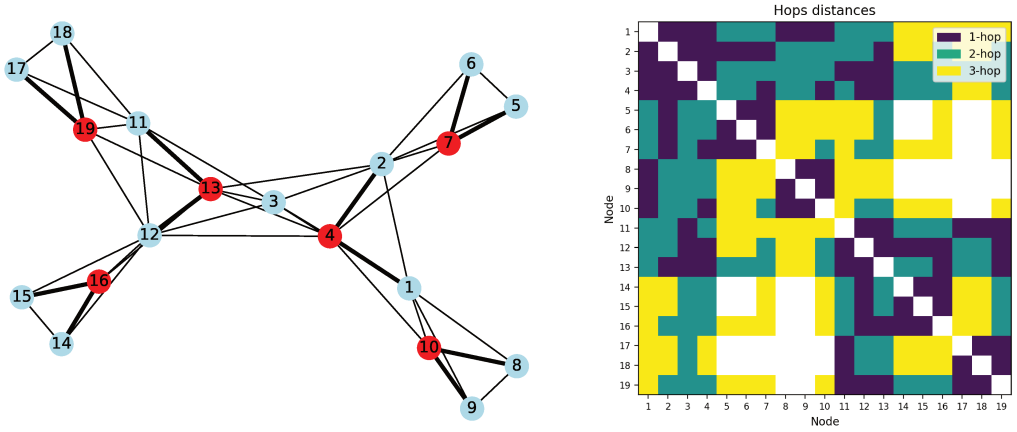


Figure 6.9: **Molecular graph and hop-distance between nodes.** Left panel: molecular graph of the protonated water hexamer when only atomic environment with $r_{\text{cut}} = 3$ Å are considered. The edges in bold are actual covalent bonds. Right panel: distances between the nodes of the molecular graph. Violet squares are the elements of the adjacency matrix and are the same colored squares of the right panel of Fig. 6.1; considering more messages passing, each node becomes 'aware' of the presence of other atoms outside of the radial cutoff.

each node in an automated manner. In the case of graphs, the predicted features of their constituents are referred to as *embeddings*, as they are directly derived from the structure in which they belong to.

While one could apply separated neural networks to each element of the graph, doing so one would miss the advantages of a graph-based representation of molecular data. Instead of considering individual entities in isolation, graph neural networks (Graph neural network (GNN)s) are a family of neural networks that can extract and use features from the underlying graph through an iterative process. To describe this node-representation learning, we adopt the framework of Message-passing neural network (MPNN) [315], a type of GNN that includes convolutional graph neural networks (CGNNs) in which different elements of the graphs exchange messages to determine their own features.

Consider a node v and its learnable features $\mathbf{h}_v^{(0)}$. These features are tuned to $\mathbf{h}_v^{(t+1)}$ at each step $t \in (0, T-1)$ through a learnable update function U_t, which depends on the previous representation, $\mathbf{h}_v^{(t)}$ and a message $\mathbf{m}_{\mathcal{N}(v)}^{(t)}$ that collects the information from its neighbouring nodes, $\mathcal{N}_v$.

$$\mathbf{h}_v^{(t+1)} = U_t\left(\mathbf{h}_v^{(t)}, \mathbf{m}_{\mathcal{N}(v)}^{(t)}\right). \tag{6.103}$$

Here, each iteration can be thought as analogous to a layer in standard neural network, which is why we used the same indexing (t) for the iterative step and the layers of a NN in Sec. 6.5. The message $\mathbf{m}_{\mathcal{N}(v)}^{(t)}$ is constructed by a *pooling* function ρ, designed to be permutationally invariant. Specifically, m involves of two operations: *gather* the neighbouring embeddings through single messages function $\mathbf{M}(\mathbf{h}_w)$, and *aggregate* them with a permutationally invariant operation, the

simplest being the summation.

$$\mathbf{m}_{\mathcal{N}(v)}^{(t)} = \rho\left(\left\{\mathbf{M}(\mathbf{h}_w^{(t)})\right\}_{w\in\mathcal{N}(v)}\right) \tag{6.104}$$

Once all the features have been computed, a learnable readout function f, which acts on all the nodes, either separately or collectively, produces the prediction we are interested in

$$\mathbf{y}^* = f\left(\left\{\mathbf{h}_v^{(T)}\right\}_{v\in V}\right) \tag{6.105}$$

In the MPNN framework, all the components described as *learnable* can be implemented using a neural network. Graph neural network are a rapidly expanding field of research with applications to network science. For a concise yet comprehensive reference on GNN see [316]. In the following we will focus on how the above framework can be exploited in machine learning chemistry.

Quantum chemistry simulation and drug discovery have been major drivers for GNN development, with numerous MPNN models emerging over he past decade [317]. Message-passing offers the appealing feature of going beyond the local atomic environment approximation by allowing multiple messages to be exchanged across the molecular graph. In the right panel of Fig. 6.9, we see a generalized adjacency matrix. The adjacency matrix $\mathbf{A} \in \mathbb{R}^{|V|\times|V|}$ encodes the connectivity of the molecular graph, indicating the presence of edges:

$$\begin{cases}\mathbf{A}[u,v] = 1 & \text{if} \quad (u,v)\in E \\ \mathbf{A}[u,v] = 0 & \text{otherwise}\end{cases} . \tag{6.106}$$

With the first messages (violet squares), the nodes know their local atomic environment, like the local descriptors we introduced in 6.2. With subsequent updates of the embeddings (green and yellow squares respectively), the nodes can receive information about atoms beyond the atomic environment radial cutoff. Of course the updates come with an non-neglible computational cost. Additionally, the internal features must be equivariant, meaning that the messages must encode the necessary isometries, much like how convolution in convolutional neural network encodes the translational symmetry.

One of the most promising MPNN-potential is MACE [318], based on ACE [282] which extends beyond the local description of the atomic environment. Each atom/node v in the layer (t) is represented by a state

$$\sigma_i^{(t)} = \left(\mathbf{q}_v, Z_v, \mathbf{h}_v^{(t)}\right) \tag{6.107}$$

where the position $\mathbf{q}_i$ and the chemical element Z_v do not change at each update, while it is the case of the node embeddings $\mathbf{h}_v^{(t)}$, which contains equivariant tensors. The embeddings are updated following the scheme of typical MPNN, using a linear combination of the messages coming from neighbouring atoms. The messages are constructed by embedding the edges using learnable radial basis and spherical harmonics containing the angular information of the neighbours, in combination with previous node features. In order guarantee features equivariance, the 2-body features pooled from the neighbouring atoms and combined through Clebsch-Gordan coefficients. Higher order features are built using the same procedure, with the difference that they are combined through tensor products, which are then symmetrised.

6.6 Machine learning potentials for neutral and protonated water

Apart from the already cited ML-based MBE-potentials [244], several MLIPs for water have been proposed and reviewed [319]. Kernel-based and HDNN potentials are the most used MLIPs, both showing similar performance, with the most critical step being database construction [320]. HDNN fitted to DFT data enabled studies on the influence of van der Waals interactions on hydrogen-bond structures [321]. Solid and liquid water have been simulated using MACE-based foundation models [322] and DeePMD, a deep neural network potential with automatic representation learning [323].

Nearly all the simulations mentioned so far are DFT-based, inheriting its limitations, particularly the choice of the exchange-correlation functional and water overstructuration. At the same computational cost, higher accuracy datasets from explicitly correlated methods are necessarily smaller in size, leading to methods that focus on refining a base model. This can be achieved by initially training on a dataset computed with lower-tier electronic structure methods, such as DFT or MP2, followed by two possible approaches: building a model that learns the difference with a more complex method like CCSD(T) (a technique known as Δ-learning) [324]), or fine-tuning the model itself through transfer learning.Transfer learning involves improving the weights of an already trained model by learning from a few high-accuracy energies. This as been demonstrated also for bulk water, progressing from HF, BLYP, revPBE0-D3 to CCSD, CCSD(T), and auxiliary-field Monte Carlo [325]. Additionally, handling long-range electrostatics with a simpler model allows the MLIP to focus on short-range interactions, which can also be seen as stacking models of increasing accuracy, as done in combining interpolated multipoles with ML-learned short-range interactions in flexible cartesian multiple combined with GAP (FCM/GAP) [326]. These techniques have been applied to bulk water, as demonstrated by augmenting a simple electrostatics model based on partial charges with a HDNN trained via transfer learning from DFT-level to MP2 and then CCSD(T)-level [327].

Managing long-range interactions remains one of the hardest tasks in MLIPs [274, 328], a longstanding problem also noted in previous classical water force fields, as briefly reviewed in Chap. 5. Long Distance Equivariant (LODE) descriptors have been proposed to capture long-range interactions based on the local value of an atom-density potential [329]. Additionally, MLIPs that consider polarization and charge transfer [330] are still under development, though most have not yet been applied to water systems.

The only type of MLIPs of this kind applied to water systems use the local atomic environment of a single molecule to infer the position of maximally localized Wannier centers (MLWCs), computed from DFT. This approach has the benefit of relying solely on *ab initio* calculations without the need for an arbitrary definition of partial charges. Notable examples include the self-consistent neural network (SCFNN) by Remsing et al. [331, 332], and the DeePMD extension with electrostatics by Car et al. [333, 334], which has been successfully applied to water ionization [335, 336]. Extending these models to charged systems or other methods beyond DFT would be interesting, though this generalization is not straightforward.

Simulations of protonated water using MLIPs have been more limited in terms of methods and systems compared to neutral water [319]. Notable studies include HDNN trained on DFT data [337], extending up to the protonated water octamer, $H^+(H_2O)_8$, and HDNN trained on high-accuracy CCSD(T) for the Zundel cation [338], as well as the hydronium ion [339] and the Eigen complex [340]. Learning multiple clusters $H^+(H_2O)_{n=1,\cdots,4}$ [341] enabled extrapolation to the protonated water hexamer in its Zundel configuration [342].

We are not aware of works applying MLIPs directly to the protonated water hexamer treated with advanced methods able to taking into account electronic correlation, such as CCSD(T) or QMC.

Chapter 7

Assessing the quality of MLIPs trained on stochastic datasets

In the last two decades machine learning interatomic potentials (MLIPs) like those presented in Chapter 6 emerged as a tool to combine the speed of parametrized potentials with the accuracy of sophisticated electronic structure methods, bridging the best of the two worlds [275]. By replacing application-tailored functional forms of typical force fields (Chapter 5) with a data-driven approach, MLIPs can fit any PES, provided that a large enough set of single-point calculations done with any electronic structure technique is available. Yet, they are mainly used to fit energy and forces that might be biased by the underlying approximations, or by the level of theory. Therefore, we find appealing employing them with QMC estimates of the PES, which are very accurate and, despite the noise, unbiased. This approach has already been successfully applied in several studies [325, 343–345], eventually in combination with Δ-machine learning [346]. The effect of noise on the learning algorithms has also been investigated in some previous works [347–349].

In this Chapter we undertake a thorough study on the robustness of MLIPs in learning noisy PES estimated with stochastic electronic structure methods. One of our goals is to answer to the following questions: how does the QMC noise affect the quality of the simulations? What is the "breaking point" of MLIPs with respect to the noise amplitude? How is this related to the size of the training set? As pointed out in Ref. [348], the trade-off between the number of datapoints used in the training, and the accuracy of each estimate, meant as stochastic error on the single datapoint, is one of the keys to efficiently exploit QMC methods in the context of MLIPs. However, it is not clear up to which level of noise this trade-off can be applicable. In order to study the learning efficiency as a function of a progressively larger noise level, we purposedly corrupted a model PES with gradually increasing noise, on which different types of MLIPs were trained. To quantify their reliability we then analyzed not only the corresponding standard test errors and learning curves, but also we carried out production runs to measure the standardized difference of physical observables between *ab initio* dynamics and MLIP-driven one.

In Section 7.1 we introduce Ceperley's classification of the noise and errors involved when applying fitting methods to data affected by noise. This scheme allows one to design a learning protocol and systematically interpret the results and the performance of the learning algorithms.

In the same spirit of the first part of this thesis, we want to study protonated water clusters. Therefore, in Section 7.2 we introduce the benchmark of our choice for the application of MLIPs, namely the Zundel ion, $H_5O_2^+$. Being the smallest protonated water cluster, this system is the starting point to study the impact of NQEs and the mechanism of proton transfer in water. Indeed, it requires an explicit quantum treatement of the nuclei to properly account for all its features. Thus the Zundel ion is a good benchmark to test the reliability of MLIPs in reproducing proton hopping between water molecules, and their robustness in RPMD simulations.

In Section 7.3 we briefly summarise the datasets on which this work is based. All our protonated water clusters datasets are sampled by Langevin dynamics (LD), a flavour of MD where the NVT ensemble is sampled using a stochastic thermostat; both its classical and path integral variants are introduced in Chapter 3. The PES reference is provided by a stochastic method, precisely variational Monte Carlo (Chapter 2), and a deterministic one based on the Many-body expansion (MBE), where the n-body terms are fitted to energies from deterministic computational chemistry methods, as described in Chapter 5.

This latter dataset is then corrupted with noise of increasing intensity aimed to imitate the effects of QMC stochastic sampling of energy and forces, as discussed in Section 7.4. While reducing the Gaussian white noise has long been a key focus in the QMC community, the structure of noise across samples of the PES has been less explored. However, this issue can potentially play a critical role in the context of PES fitting with MLIPs.

Once the QMC noise is correctly reproduced and added to the clean MBE energy and forces, we train on such datasets both kernel- and neural network-based MLIPs, following the protocol described in Section 7.5.

In Section 7.6 we outline our comparative approach: while standard tests errors and learning curves represent the most direct way to probe the MLIPs, as we do in Section 7.6.1, the importance of tests based on actual physical quantities in assessing the quality of a MLIP has been demonstrated [350, 351]. For this reason we included in the analysis the evaluation of both static quantities, like the radial distribution functions, and dynamic quantities, like the velocity autocorrelation functions. This evaluation is carried out by averaging the standardized difference between the above physical quantities computed along different trajectories initialized with different starting configurations, as explained in Section 7.6.2.

Finally, in Section 7.7 we show some preliminary results on the learning of the protonated water hexamer.

7.1 Classification of noise and errors

We are interested in finding the true energy $E(\mathbf{q}_j)$ and the true forces $\mathbf{f}(\mathbf{q}_j)$ of each configuration belonging to a dataset of $N = N_{\text{test}}$ configurations, $\{\mathbf{q}_j\}_{j=1,\cdots,N_{\text{test}}}$, using a MLIP trained on N_{train} configurations, $\{\mathbf{q}_i, E_i, \mathbf{f}_i\}_{i=1,\cdots,N_{\text{train}}}$, where $E_i = E(\mathbf{q}_i)$ and $\mathbf{f}_i = \mathbf{f}(\mathbf{q}_i)$.

We follow the work of Ceperley et al. [348] in the definition of the errors involved when training MLIPs, as represented in Fig. 7.1.

Dataset stochastic error δ and standard deviation σ

Let $E_i^{\text{V}} = E_{\text{V}}(\mathbf{q}_i)$ and $\mathbf{f}_i^{\text{V}} = \mathbf{f}_{\text{V}}(\mathbf{q}_i)$ be the VMC estimates of the energies and the forces, respectively. For the sake of readability, we will put a v in superscript whenever a quantity is estimated by a single-point VMC run[1].

The total error in E_i^{V} and $\mathbf{f}_i^{\text{V}}$ consists of two components: the *stochastic error*, δ_{E_i} and $\boldsymbol{\delta}_{\mathbf{f}_i}$, which arises from the intrinsic randomness of the quantum Monte Carlo method, and the *systematic error*, or bias, which results from the approximations inherent in the method itself, like the basis set error. Hereafter, we assume that the VMC estimates E_i^{V} and $\mathbf{f}_i^{\text{V}}$ are unbiased, namely they are not affected by any systematic error.

In this study we also assume that the true energies and forces are known. This is in general not true, but in our "whole dataset fitting" approach, the knowledge of the ground truth will be used to study the robustness of the ML process against noise. Then, the stochastic error associated with each element of the dataset is known and can be written as

$$\begin{aligned} \delta_E^{\text{V}}(\mathbf{q}_i) &= E_i - E_i^{\text{V}}, \\ \boldsymbol{\delta}_{\mathbf{f}}^{\text{V}}(\mathbf{q}_i) &= \mathbf{f}_i - \mathbf{f}_i^{\text{V}} = \mathbf{f}_i - \nabla_{\mathbf{q}_i} E_i^{\text{V}}. \end{aligned} \tag{7.1}$$

where in the case of the forces we have a vector of stochastic errors, one for each component:

$$\boldsymbol{\delta}_{\mathbf{f}}^{\text{V}} = \left(\delta_{f,ia_1x}^{\text{V}}, \delta_{f,ia_1y}^{\text{V}}, \delta_{f,ia_1z}^{\text{V}}, \delta_{f,ia_2x}^{\text{V}}, \cdots, \delta_{f,ia_Mx}^{\text{V}}, \delta_{f,ia_My}^{\text{V}}, \delta_{f,ia_Mz}^{\text{V}} \right). \tag{7.2}$$

We define the *stochastic error vectors associated to the dataset*, $\boldsymbol{\delta}_E$ and $\boldsymbol{\delta}_{\mathbf{f}}$, as those vectors whose entries are QMC errors associated to each single configurations in the whole test set:

$$\boldsymbol{\delta}_E = \begin{pmatrix} \delta_{E_1}^{\text{V}} \\ \delta_{E_2}^{\text{V}} \\ \vdots \\ \delta_{E_N}^{\text{V}} \end{pmatrix} = \begin{pmatrix} E_1 - E_1^{\text{V}} \\ E_2 - E_2^{\text{V}} \\ \vdots \\ E_N - E_N^{\text{V}} \end{pmatrix} \tag{7.3}$$

[1] In the notation of Chapter 2, we would have $\overline{E}(\mathbf{q}) = E^{\text{V}}(\mathbf{q})$ and $\bar{\mathbf{f}}(\mathbf{q}) = \mathbf{f}^{\text{V}}(\mathbf{q})$.

$$\boldsymbol{\delta}_{\mathbf{f}} = \begin{pmatrix} \delta^{\mathrm{V}}_{f_{1,1,x}} \\ \delta^{\mathrm{V}}_{f_{1,1,y}} \\ \delta^{\mathrm{V}}_{f_{1,1,z}} \\ \vdots \\ \delta^{\mathrm{V}}_{f_{N,M,x}} \\ \delta^{\mathrm{V}}_{f_{N,M,y}} \\ \delta^{\mathrm{V}}_{f_{N,M,z}} \end{pmatrix} = \begin{pmatrix} f_{1,1,x} - f^{\mathrm{V}}_{1,1,x} \\ f_{1,1,y} - f^{\mathrm{V}}_{1,1,y} \\ f_{1,1,z} - f^{\mathrm{V}}_{1,1,z} \\ \vdots \\ f_{N,M,x} - f^{\mathrm{V}}_{N,M,x} \\ f_{N,M,y} - f^{\mathrm{V}}_{N,M,y} \\ f_{N,M,z} - f^{\mathrm{V}}_{N,M,z} \end{pmatrix} = \begin{pmatrix} f_{1,1,x} - \partial E^{\mathrm{V}}_1 / \partial q_{1,x} \\ f_{1,1,y} - \partial E^{\mathrm{V}}_1 / \partial q_{1,y} \\ f_{1,1,z} - \partial E^{\mathrm{V}}_1 / \partial q_{1,z} \\ \vdots \\ f_{N,M,x} - \partial E^{\mathrm{V}}_N / \partial q_{M,x} \\ f_{N,M,y} - \partial E^{\mathrm{V}}_N / \partial q_{M,y} \\ f_{N,M,z} - \partial E^{\mathrm{V}}_N / \partial q_{M,z} \end{pmatrix} \tag{7.4}$$

where in the case of energies we have a N_{test}-long vector, while for forces the total length depends not only on the number of configurations in the test set, but also on the number of atoms in each configuration and the dimensionality of the coordinates. In the case of a homogeneous dataset where only a single type of systems appears in different 3D configurations, this translates in a $3 \times N_{\text{test}} \times M$-long stochastic forces error, where M is the number of atoms in the system.

We recall that in the usual QMC setting, the exact values of $\delta^{\mathrm{V}}_E(\mathbf{q}_i)$ and $\boldsymbol{\delta}^{\mathrm{V}}_{\mathbf{f}}(\mathbf{q}_i)$ are not known. They are random variables normally distributed around zero (due to the fact that we assumed unbiased VMC estimates) with variance $(\sigma^{\mathrm{V}}_E(\mathbf{q}_i))^2$ and $(\boldsymbol{\sigma}^{\mathrm{V}}_{\mathbf{f}}(\mathbf{q}_i))^2$, as estimated according to statistical methods of Chapter 2. For example, consider a single configuration in the dataset, $\mathbf{q}_i$, and its energy E_i. The QMC estimate of E^{V}_i is affected by a stochastic error with standard deviation formally defined as

$$\sigma^{\mathrm{V}}_E(\mathbf{q}_i) = \sigma[\overline{E}(\mathbf{q}_i)] = \sqrt{\frac{\mathrm{var}[E_L(\mathbf{q}_i)]}{N_{\text{gen}}}}, \tag{7.5}$$

where we made explicit the parametric dependence on the fixed nuclear configuration $\mathbf{q}_i$, for which we sampled N_{gen} electronic configurations.

As we did for the stochastic errors, we can define the dataset *standard deviation vectors*, $\boldsymbol{\sigma}_E$ and $\boldsymbol{\sigma}_{\mathbf{f}}$, whose "dataset norm" is given by

$$|\boldsymbol{\sigma}_E| = \sigma_E = \sqrt{\frac{1}{N}\sum_i^N \left(\sigma^{\mathrm{V}}_{E_i}\right)^2} \qquad \text{and} \qquad |\boldsymbol{\sigma}_{\mathbf{f}}| = \sigma_{\mathbf{f}} = \sqrt{\frac{1}{3MN}\sum_i^N \sum_a^M \sum_d^{x,y,z} \left(\sigma^{\mathrm{V}}_{f_{i,a,d}}\right)^2}, \tag{7.6}$$

which are an estimate of the dataset norms of $\boldsymbol{\delta}_E$ and $\boldsymbol{\delta}_{\mathbf{f}}$.

Fitting error ρ

In practice, only the QMC estimates E^{V} and $\mathbf{f}^{\mathrm{V}}$, are really available, so when we measure the *fitting error* $\boldsymbol{\rho}_E$, also defined as *test error* in the ML context, we are dealing with the following quantity:

$$\boldsymbol{\rho}_E = \begin{pmatrix} E^{\mathrm{V}}_1 - E^m_1 \\ E^{\mathrm{V}}_2 - E^m_2 \\ \vdots \\ E^{\mathrm{V}}_N - E^m_N \end{pmatrix} \qquad \rho_E = \sqrt{\frac{1}{N}\sum_i^N \left[E^{\mathrm{V}}_i - E^m_i\right]^2}, \tag{7.7}$$

where we used the root mean square error (RMSE), at variance with the mean absolute error (MAE), which is also used in ML. An analogous definition holds for the forces fitting error, $\boldsymbol{\rho}_{\mathbf{f}}$.

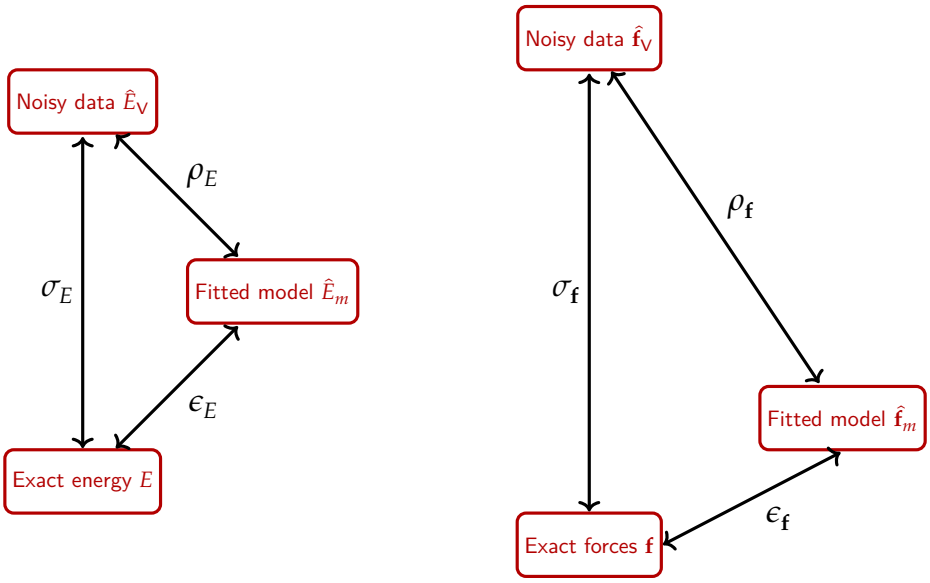


Figure 7.1: **Relationship between the true observables, their stochastic estimate and their machine learning model prediction trough the three types of errors: σ, ρ, ϵ.** Fixing the number of test configurations, N_{test}, and defining the errors as vectors allows one to relate them through triangular inequalities: $|\boldsymbol{\rho} - \boldsymbol{\sigma}| \leq \boldsymbol{\epsilon} \leq |\boldsymbol{\rho} + \boldsymbol{\sigma}|$. From [348].

In noiseless datasets, the test error is expected to decrease with an increasing training set N_{train} following a power law [289]

$$\rho \sim \mathcal{O}(N_{\text{train}}^{-\alpha}), \tag{7.8}$$

which defines the "learning curve". For noisy dataset it is expected that a plateau dependent on the average noise will limit ρ from below, such that:

$$\rho \sim \mathcal{O}(N_{\text{train}}^{-\alpha}) + g(\sigma). \tag{7.9}$$

Model error ϵ

On a more regular basis, what we are interested in is the *model error*, that is, how far the model is from the ground truth values:

$$\boldsymbol{\epsilon}_E = \begin{pmatrix} E_1^m - E_1 \\ E_2^m - E_2 \\ \vdots \\ E_N^m - E_N \end{pmatrix} \qquad \epsilon_E = \sqrt{\frac{1}{N}\sum_i \left[E_i^m - E_i\right]^2}. \tag{7.10}$$

Unfortunately the only error available during the training and test procedure is the fitting error, so it would be interesting to study the relation between the three types of error on E and $\mathbf{f}$, $\{\sigma_E, \rho_E, \epsilon_E, \sigma_\mathbf{f}, \rho_\mathbf{f}, \epsilon_\mathbf{f}\}$, exemplified in Fig. 7.1, in a controlled setup. In the perspective of applying MLIPs to study proton transfer in water, we simulate small protonated water clusters using deterministic PES, gradually adding stochastic noise to the clean dataset. The deterministic PES will be our ground truth and the dataset corrupted by adding stochastic noise will mimic

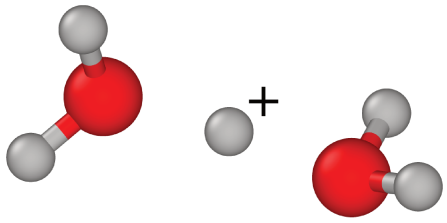


Figure 7.2: **The Zundel cation.**

the behaviour of a QMC dataset with a tunable stochastic error. As introduced in Chapter 5, among the best and efficient methods to study protonated water clusters there are the many-body expansion-based PES, which are complex enough to reproduce some non-trivial aspects of the proton in water. The simplest of such systems is the Zundel ion, which is the topic of the next Section.

7.2 The Zundel ion

The Zundel ion [18], $H_5O_2^+$ in Figure 7.2, is the smallest water cluster exhibiting non-trivial proton transfer, and its compact size facilitates a comprehensive and systematic study of the PT problem.

Indeed the Zundel ion has been extensively studied with all possible electronic structure methods, from DFT [352] and CPMD [162, 353], to MP2 [354, 355] and coupled cluster including different levels of excitations, CCD(T) [356] and CCSD(T) [355, 357]. Furthermore, due to the significant influence of NQEs in hydrogen dynamics, the Zundel cation is frequently used to test and validate new approaches able to deal with quantum nuclei, such as multiconfiguration time-dependent Hartree [52]), multiple time step integrators and ring-polymer contraction [358], and the PIOUD algorithm itself [133].

More importantly, a vast amount of experimental data is available for comparison. The rapid development of spectroscopic instruments has enabled the probing of vibrational properties in ionic species, leading to numerous studies [194, 359] on the $H_5O_2^+$ ion.

Given the availability of large quantity of experimental and theoretical data, the Zundel is widely used as a benchmark system for parametrized PES and new *ab initio* methods as well. Within the QMC framework, the accuracy of the Jastrow correlated AGP wave function has been tested on the Zundel complex [130]. Another class of methods tested on this system are those based on MS-EVB [263–265, 360], which paved the way for their application to extended systems [361, 362]. At the same time, parametrised PES generated through ML schemes [338] have also been benchmarked on the Zundel cation. We recall also that among the first applications of PIPs-fitted PES at coupled cluster accuracy, there is the one from Bowman [255], based on the many-body expansion (Chapter 5). This is the PES that we will use as deterministic reference model

7.3 Datasets description

The datasets are generated using Langevin dynamics because stochastic thermostatting is well-suited to deal with both noisy forces, such as those coming from QMC, and deterministic ones, allowing for a direct comparison between QMC and MLIP trajectories. For classical simulations over deterministic PES we use the Bussi algorithm (Section 3.2.3); for classical simulation over stochastically estimated PES we use the Attaccalite-Sorella algorithm (Section 3.2.4); for quantum simulations we use the PIOUD algorithm (Section 3.5.2), as it is designed to run with both deterministic and stochastic forces. The Zundel dynamics are driven by the deterministic MBE-PES [255]. We generate trajectories at different temperatures from 50 K to 600 K, comprising 30.000 steps of $\delta t = 0.5$ fs, for a total of 15 ps of physical simulation time.

In Figure 7.3 we plot some of the MD-generated datasets in a space of reduced dimension. Specifically we employed the principal covariates regression (PCovR) technique [363], a combination of principal component analysis (PCA) and linear regression which allows one to visualize basic structure-property relationships. In our case we applied it to local atomic environments of the Zundel obtained from the 300K MBE-driven trajectory, which will also be our starting point for the application of MLIPs. In Fig.7.3a we see that, by considering the total energy as a target, all configurations are sorted for increasing energy values along PCovR[1]. At the same time, in Fig.7.3 we notice that the second principal covariates correlates well with the oxygen-oxygen distance, without explicit human input. This distance is one of the most important internal coordinates in the Zundel cation, as it is for the protonated water hexamer as well (Chapter 4). The automatic recognition of d_{OO} as one of the principal covariates indicates both good sampling of the configuration space and the appropriateness of the dimensionality reduction parameters.

Once we can rely on this technique, we can use it to visualize the other datasets. The bottom panels of Fig. 7.3 illustrates how increasing the temperature translates to a wider exploration of the PES, especially in classical simulation. On the other hand, we see that the inclusion of NQEs through RPMD makes the necklace explore a larger space, and this happens already at low temperatures, where a higher number of beads can bring the ring polymer configurations far from the classical ones.

We also use two classical trajectories based on VMC-estimated PES, one at 50 K of 39999 steps and another at 300 K of 19163 steps. The former is much longer because at each step of the dynamics the electronic QMC sample to estimate $E^{\mathrm{V}}(\mathbf{q})$ and $\mathbf{f}^{\mathrm{V}}(\mathbf{q})$ is smaller: $N_{\mathrm{gen}} = 81920$ against $N_{\mathrm{gen}} = 331776$.

Besides these datasets extracted directly from (PI)MD simulations, we generated noisy datasets by adding artificial stochastic errors to the deterministic baseline (see Section 7.4).

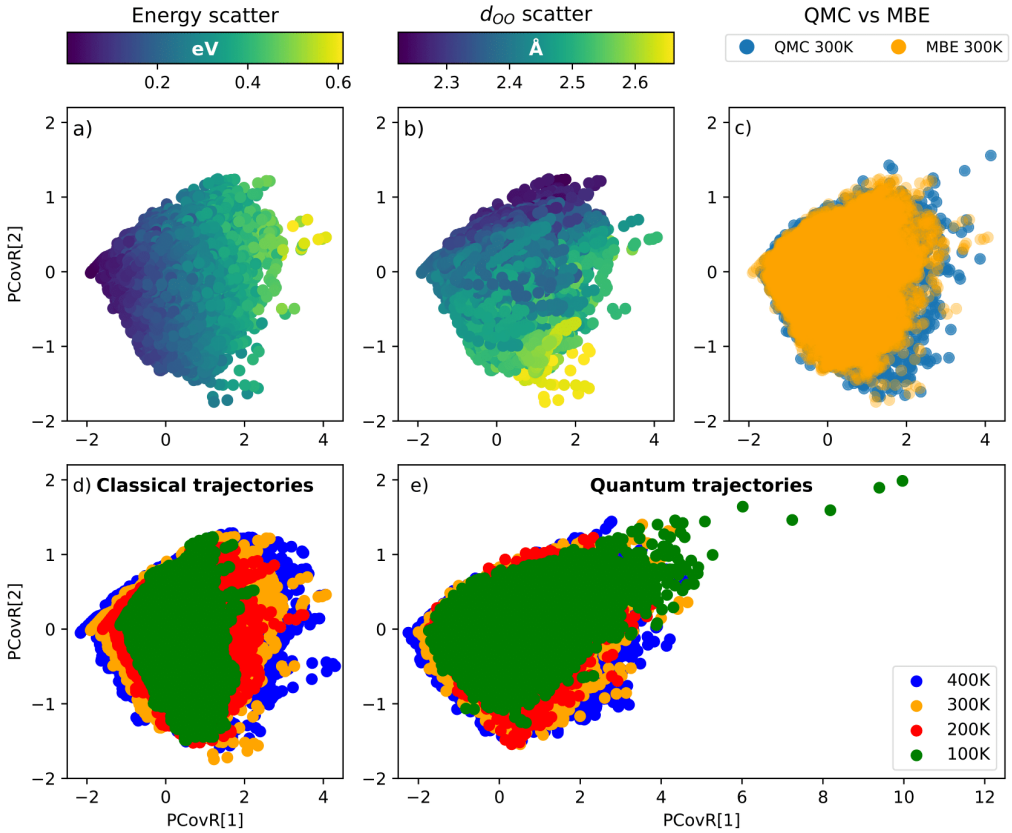


Figure 7.3: **Dimensionality reduction of the Zundel ion classical and quantum trajectories.** The dimensionality reduction algorithm is based on the SOAP representation of the local atomic environments, as implemented in the `Librascal` package [364]. SOAP is conceptually similar to the ACE descriptor presented in Chapter 6. The SOAP parameters are the following: cutoff=3.0, max_radial=6, max_angular=4, atomic_gaussian_width=0.3, cutoff_function='ShiftedCosine' of width=0.5, radial_basis='Gto' with accuracy=1e-6, center_atom_weight=1.0.

7.4 Applying Gaussian noise to energies and forces

Given a configuration $\mathbf{q}$, the QMC estimates of its energy and forces are respectively affected by uni-variate and multivariate Gaussian white noise. These noises are characterized by the standard deviations $\sigma_E^{\mathrm{V}}(\mathbf{q})$ and $\sigma_{\mathbf{f}}^{\mathrm{V}}(\mathbf{q})$, which depend on the number of QMC stochastic samplings N_{gen} as $1/\sqrt{N_{\mathrm{gen}}}$. When several configurations $\{\mathbf{q}_i\}_{i=1,\cdots,N_{\mathrm{MD}}}$ are collected along the trajectory, and for each of them E and $\mathbf{f}$ are estimated with the same number of QMC samples N_{gen}, it is not a priori clear how the standard deviations $\sigma_E^{\mathrm{V}}(\mathbf{q}_i)$ and $\sigma_{\mathbf{f}}^{\mathrm{V}}(\mathbf{q}_i)$ are distributed across the training dataset. In other words, we do not know whether different points on the PES have the same error bars, a condition known as *homoscedasticity*, or if they have varying error bars, referred to as *heteroscedasticity*. We are aware only that pointwise the energy error of a single configuration should follow a normal distribution.

To analyse their actual distribution we take all the configurations and their respective standard deviations, $\{\mathbf{q}_i, \sigma_E^{\mathrm{V}}(\mathbf{q}_i), \boldsymbol{\sigma}_{\mathbf{f}}^{\mathrm{V}}(\mathbf{q}_i)\}$, from classical QMC-driven MD simulations for both $H_5O_2^+$ and $H_{13}O_6^+$ (Chapter 4). By binning $\sigma_E^{\mathrm{V}}(\mathbf{q}_i)$ for both systems in Figures 7.4a and 7.4c, it turns out that its distribution is almost normally distributed. However, this does not seem to be the case for the standard deviation in the force components, σ_f^{V}, even when selecting the contributions from a specific species (Figure 7.4b and 7.4d). This behaviour is partially explained by the finite size of the system in open boundary condition, at variance with periodic systems where the average force experienced by the atoms is more isotropic. Another source of modulation comes from the non-equivalent role played by the different ions in the system. The relative size of the errors will thus depend on the corresponding force component. Figure 7.4h, relative to the protonated water hexamer, clearly shows that atoms of the same species but having different roles in the cluster can show different distributions in the error. Specifically, the oxygen atoms in the Zundel core exhibit larger errors compared to those in the solvation shell (see the two red peaks), and the central proton is affected by a larger error than all the other hydrogens (the green and orange peaks, respectively). On the other hand, the multi-modal distribution of $\boldsymbol{\sigma}_{\mathbf{f}}$ components can be mapped into a single-peaked one when we consider the norm $|\boldsymbol{\sigma}_{\mathbf{f}}|$ of the 3M-dimensional vectors, as in Figures 7.4e and 7.4g. σ_E and $|\boldsymbol{\sigma}_{\mathbf{f}}|$ are single-peaked because they represent collective properties of the system as a whole.

Although computing the norm of the entire forces error vector restores the isotropy of the error distribution, from species-selected plots (Figures 7.4f and 7.4h) it is apparent that in order to mimic the QMC error in an inhomogeneous and finite-size system it is necessary to consider different standard deviations for each species belonging to the system. The qualitative analysis above give us some indication on how to produce the artificial noise with which we will corrupt the deterministic datasets. In the case of energies, the standard deviation $\hat{\sigma}_E$ is taken as the average over all the ensemble $\{\sigma_E(\mathbf{q}_k)\}_{k=1,\cdots,N_{\mathrm{tot}}}$ in the dataset. Regarding the forces, we average over all the forces standard deviation components $\sigma_f(\mathrm{X})$ affecting a specific species X, including their multiplicity M_{X}:

$$\sigma_{f_x}(\mathrm{X}) = \sigma_{f_y}(\mathrm{X}) = \sigma_{f_z}(\mathrm{X}) = \sigma_f(\mathrm{X}) = \sqrt{\frac{1}{3N_{\mathrm{tot}}M_{\mathrm{X}}} \sum_{k}^{N_{\mathrm{tot}}} \sum_{a}^{M_{\mathrm{X}}} \sum_{d}^{x,y,z} \left(\sigma_{f_{i,a,d}}^{\mathrm{V}}\right)^2}. \tag{7.11}$$

The species-specific averages of the standard deviations are reported in Table 7.1, highlighting the multivariate property of the forces noise.

Pushing the standard deviation analysis further, we can study how the specific geometry of the system at hand influences the error distribution. Indeed it has been suggested [147] that the variance of the local forces $\mathbf{f}_L$ (Chapter 2) acting on a fixed configuration $\mathbf{q}$, formally defined as

$$\mathrm{var}\left[\mathbf{f}_L\right] = \mathbb{E}\left[\left(\mathbf{f}_L - \mathbf{f}_{\mathrm{VMC}}\right)\left(\mathbf{f}_L - \mathbf{f}_{\mathrm{VMC}}\right)^T\right], \tag{7.12}$$

is proportional to the dynamical matrix, that is, to the Hessian $\mathbf{H}(\mathbf{q})$, computed as the second derivative of the energy with respect to all the couples of cartesian coordinates

$$H_{xy} = \frac{\partial^2 E(\mathbf{q})}{\partial q_x \partial q_y}. \tag{7.13}$$

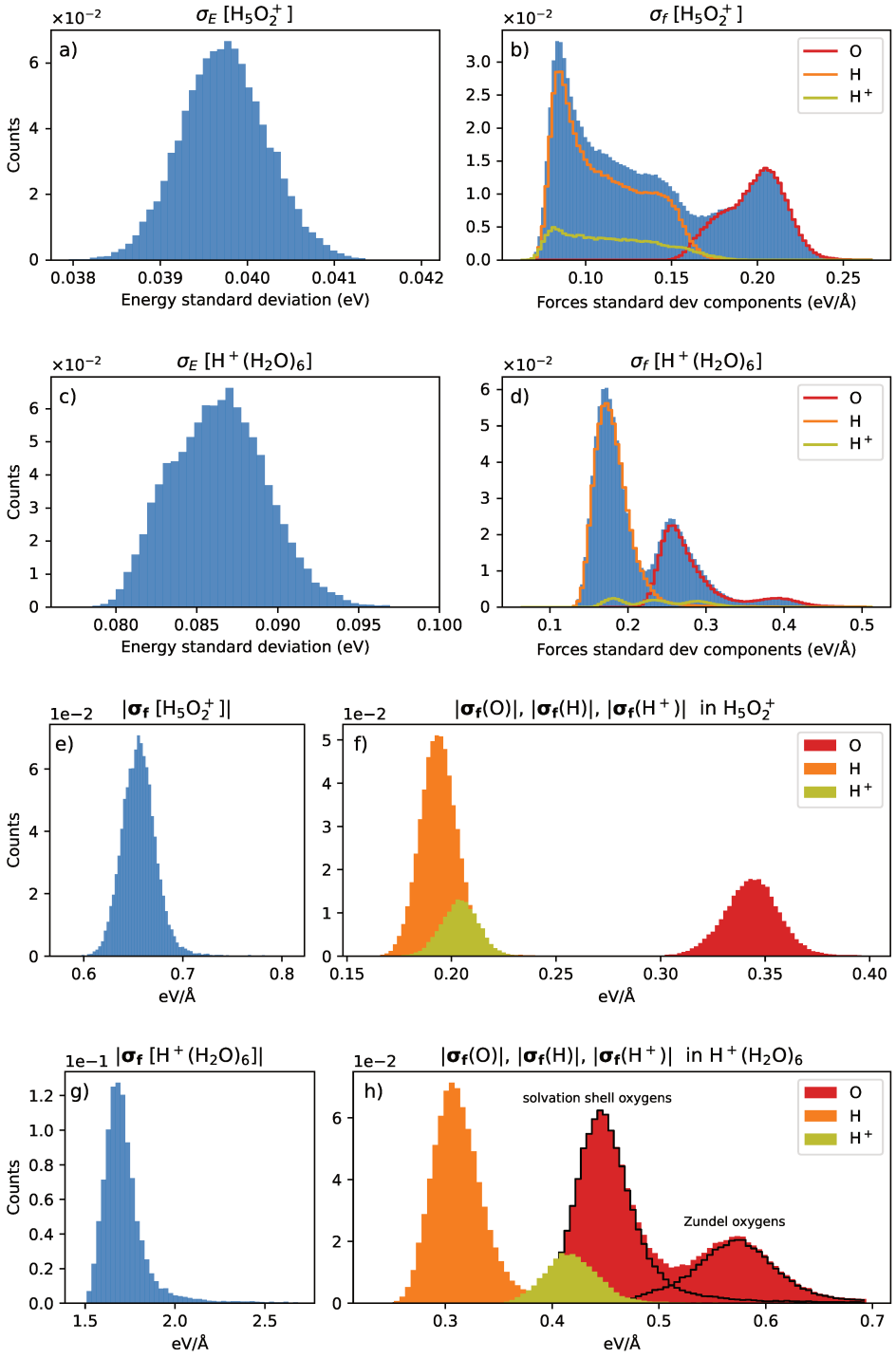


Figure 7.4: **Histograms of QMC energy and forces standard deviations in $H_5O_2^+$ and $H_{13}O_6^+$ in classical simulations at** 300 K **and** 250 K**, respectively.** While the energy standard deviation histograms (a,c) have a shape that can be compared to a Gaussian bell, the forces in both the Zundel ion and the protonated water hexamer show several peaks (b,d). These peaks can be associated to different elements, but even with this resolution their shape is far from the normal one (coloured lines in b and d). To restore a Gaussian distribution in forces standard deviations, we compute their norm (e,g). Notice that different elements are characterized by different standard deviations (f,g). This distinction is necessary even between the two oxygens in the Zundel core of the protonated water hexamer, and those that belong to the solvation shell (black lines in g).

Average standard deviation	$H_5O_2^+$
σ_E	41 meV
σ_f	134 meV/Å
$\sigma_f(\mathrm{O})$	201 meV/Å
$\sigma_f(\mathrm{H})$	108 meV/Å
$\sigma_f(\mathrm{H}^+)$	113 meV/Å

Table 7.1: **Average QMC standard deviations along the trajectory generated by a QMC-driven classical MD simulation of the Zundel ion at** 300 K**.**

Since the forces are vectorial quantities, the expression in Eq. (7.12) is called *variance-covariance matrix*, $\mathbf{\Sigma_f}(\mathbf{q})$, and it reads

$$\mathbf{\Sigma_f}(\mathbf{q}) = \begin{pmatrix} \mathrm{var}\left[f_{1x}(\mathbf{q})\right] & \mathrm{cov}\left[f_{1y}(\mathbf{q}),f_{1x}(\mathbf{q})\right] & \cdots & \mathrm{cov}\left[f_{Mz}(\mathbf{q}),f_{1x}(\mathbf{q})\right] \\ \mathrm{cov}\left[f_{1y}(\mathbf{q}),f_{1x}(\mathbf{q})\right] & \mathrm{var}\left[f_{1y}\right] & \cdots & \mathrm{cov}\left[f_{Mz}(\mathbf{q}),f_{1y}(\mathbf{q})\right] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}\left[f_{Mz}(\mathbf{q}),f_{1x}(\mathbf{q})\right] & \mathrm{cov}\left[f_{1y}(\mathbf{q}),f_{Mz}(\mathbf{q})\right] & \cdots & \mathrm{var}\left[f_{Mz}\right] \end{pmatrix}, \tag{7.14}$$

where the diagonal terms are the variances of the forces components, while the off-diagonal terms represent the covariances across different components. These covariances are simultaneous, meaning that they are computed for the same electronic configuration $\mathbf{r}$ during the QMC sampling, and do not measure the time lag-covariances.

In practice, the variance-covariance matrix can be estimated using the formula

$$\mathbf{\Sigma_f}(\mathbf{q}) \approx \frac{1}{N_{\mathrm{gen}}(N_{\mathrm{gen}}-1)} \sum_{i=1}^{N_{\mathrm{gen}}} \Big(\mathbf{f}_L(\mathbf{q},\mathbf{r}_i) - \mathbf{f}_{\mathrm{V}}(\mathbf{q})\Big)\Big(\mathbf{f}_L(\mathbf{q},\mathbf{r}_i) - \mathbf{f}_{\mathrm{V}}(\mathbf{q})\Big)^T, \tag{7.15}$$

where we made explicit the dependence of the local force on the electronic coordinates $\mathbf{r}_i$ sampled by the QMC algorithm. The diagonals terms of this expression are the squared standard deviation $\left(\sigma^{\mathrm{V}}_{f_{ad}}(\mathbf{q})\right)^2$.

We tested this hypothesis by computing the Hessian of all Zundel configurations sampled by QMC-driven MD . To do so we used the `Tapenade`[365] automatic differentiation tool, which for instance have been exploited also for fast and accurate computation of the forces in the dynamics. Then we computed the vector of sorted eigenvalues of $\mathbf{H}(\mathbf{q})$ for each configuration $\mathbf{q}$ in the dataset, and we compared them with the respective vectors of sorted eigenvalues of $\mathbf{\Sigma_f}(\mathbf{q})$, configuration wise. The comparison consisted in evaluating the degree of alignment of such vectors based on the normalised dot product, $\frac{1}{2} + \frac{1}{2}\frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$, obtaining an average score of 0.89.

Also, the distribution of the square root of the entries all the Hessians' diagonals in the dataset closely resembles the distribution of all the standard deviations on the force components (Figure 7.5a).

Since the average involved in Equation (7.15) is over electronic configurations, the only way the molecular structure dependence could possibly enter in such QMC stochastic estimate of observables would be through electron-ion coupling. For our purposes we simulated the QMC

noise using both Hessian-correlated multivariate Gaussian white noise (GWN), and uncorrelated multivariate GWN where just the variances are element-dependent (as in Table 7.1). We recall that in both cases we are in presence of time-uncorrelated noise, and that the correlation is meant between components of force error vector estimated in a single-point calculation. Since the Hessian does not change much among contiguous configurations along the trajectory, it could implicitly introduce some correlation in the QMC-driven dynamics, which however is taken in consideration in the noise-correcting Langevin dynamics scheme described in Chapter 2.

As in preliminary runs we did not observe much difference in the performance of MLIPs trained on both Hessian-correlated and uncorrelated noise, we limited the study on the latter type of noise, as it is easier to generate while still keeping the information contained in the diagonal of $\mathbf{H}$. Given the scalar standard deviation σ_E and the 21-dimensional one on the forces $\boldsymbol{\sigma}_\mathbf{f}$, we sampled as many random scalars $\{\delta_E(\mathbf{q}_i)\}_{i=1,\cdots,N_{\text{train}}}$ and random vectors $\{\boldsymbol{\delta}_\mathbf{f}(\mathbf{q}_i)\}_{i=1,\cdots,N_{\text{train}}}$ as configurations in our datasets, and multiplied them by different factors k. This is meant to reproduce QMC sampling at different values of N_{gen}, as reported in Table 7.2.

σ_E(meV)	11	22	27	41	54	67	77	94	109	133	149	163
σ_f(meV/Å)	39	78	97	145	194	238	274	335	388	475	531	581

Table 7.2: **Progressively increasing standard deviation on energies and forces used to produce the noise to add to MBE values.**

The fact that all the components of the forces vector can be multiplied by the same factor is graphically justified in the bottom panels of Figure 7.5, where we see that the shape of the standard deviation distribution does not change much for different values N_{gen} (Fig. 7.5b), especially in the main peaks which are those we could clearly associate to specific elements (see Fig.7.4b and d). Indeed they can be mapped to each other by rescaling them with the square root of their QMC sample size (Fig. 7.5c). Incidentally, the scale-invariance shown in Fig. 7.5c also demonstrates that the force error distribution is largely temperature independent.

We then corrupted the clean MBE energies and forces by simple addition of the noise, as if they were produced by a stochastic method, as follows:

$$\tilde{E}_k(\mathbf{q}) = E_{\text{MBE}}(\mathbf{q}) + k\delta_E(\mathbf{q}) \tag{7.16}$$

$$\tilde{\mathbf{f}}_k(\mathbf{q}) = \mathbf{f}_{\text{MBE}}(\mathbf{q}) + k\boldsymbol{\delta}_\mathbf{f}(\mathbf{q}), \tag{7.17}$$

with δ_E and $\boldsymbol{\delta}_\mathbf{f}$ normally distributed with zero mean and variance given by the square of the averaged standard deviation as reported in Table 7.2. We stress that for $\boldsymbol{\delta}_\mathbf{f}$ the variance is species dependent.

7.5 Choice of MLIPs and learning protocol

We choose two of the MLIPs model exposed in the Chap. 6:

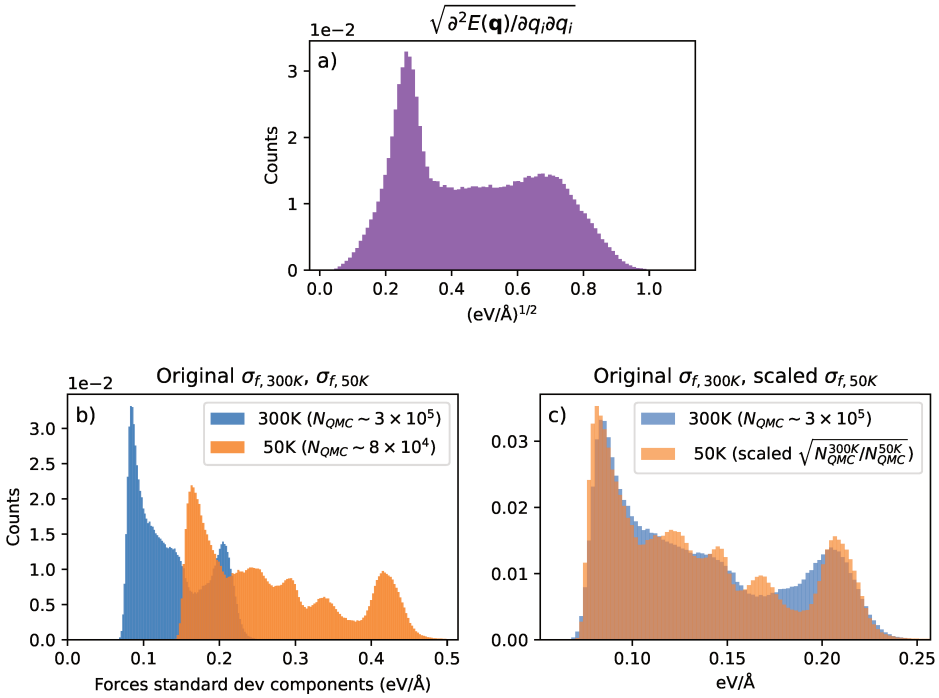


Figure 7.5: **Distribution of the square root of Hessian's diagonals entries (a) and invariance of forces standard deviation distribution for different QMC samplings (b,c) in QMC-driven classical simulations.** If the distribution of $\sigma_{\mathbf{f}}$ is affected by the geometry (the Hessian) of the molecule through the ion-electrons coupling, its multimodal shape should not change much upon different value of QMC sampling N_{gen}. This appears to be the case, as the 50K and 300K $H_5O_2^+$ datasets have been sampled with very different number of QMC step (central panel), and when we scale the 50K distribution considering the square root of the ratio of the number of steps, the two distributions overlap with a very good overlay of the main peaks (c).

- The kernel method based on operator quantum machine learning (OQML)
- The message passing neural network framework as implemented in MACE.

Some details about hyper-parameters and settings of these two methods can be found in Appendix C.

We list here the steps we followed to generated the MLIP in the OQML and MPNN framework. The procedure is standard and has been implemented using `scikit-learn` Python package [366].

Separation of training and test set with shuffling

The size of the training set is fixed so that the ratio between the N_{train} and the whole dataset size is 70%. The remaining 30% is the test set, and it is put aside during the whole training procedure. The effective training is done over smaller training sub-set of increasing size. The goal is to study the relationship in Equations (7.8) and (7.9).

Model selection and validation

Figure 7.6: **Dataset splitted into training and test sets.**

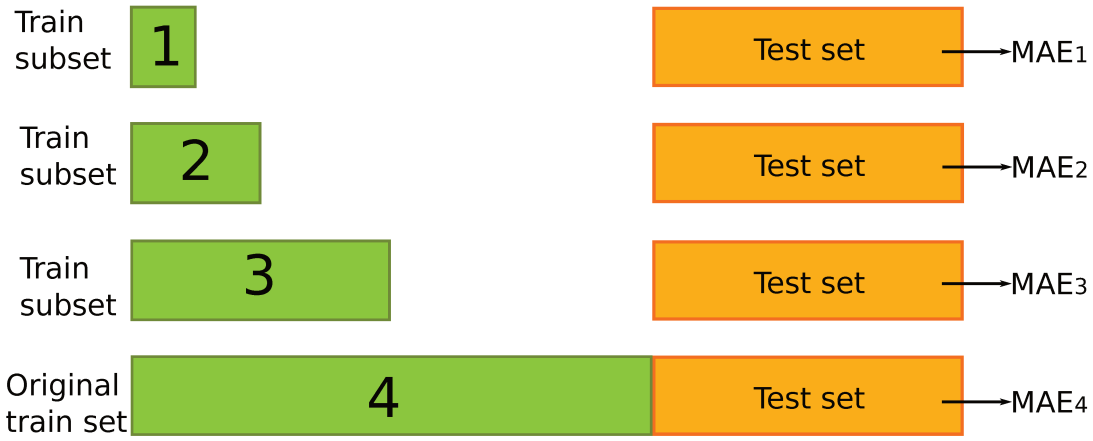


Figure 7.7: **Training set further splitted into smaller subsets of increasing size.**

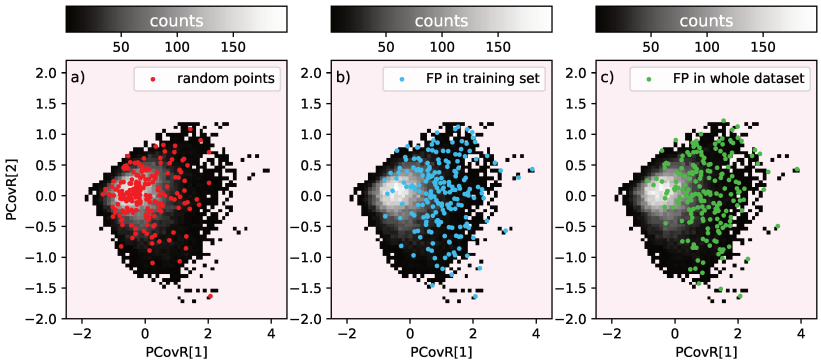
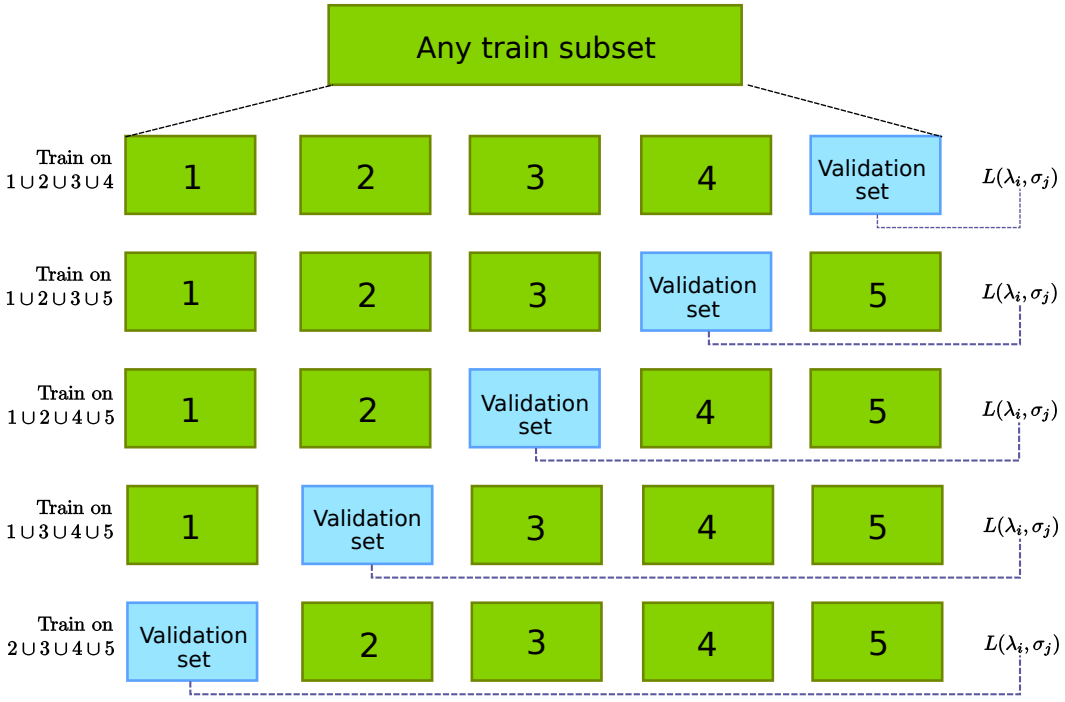


Figure 7.8: **Random sampling vs. farthest point sampling for the selection of the training subsets.** Training subsets of increasing size can be selected by different approaches. Since in this work we are mainly interested in the effect of the noise, for simplicity we will just use randomly selected configurations (a). Notice however how in this way most of the sampled configurations are concentrated in the region with an higher Boltzmann weight. While this does not constitute a big issue in the case of the Zundel, for more complex systems with different minima and transition states this could be a problem. Popular approaches include *farthest point sampling* (FPS) [367, 368], a greedy algorithm where the training points are selected by maximizing the diversity of the configurations. This diversity is commonly computed in the descriptor space (as is the case here), or it is based on a kernel-based similarity, which automatically defines a distance in the features space. One should be aware that FPS should still be applied after the first train-test split (b), otherwise the distance matrix, maximized over the whole dataset, would exploit the information of configurations not included in the training set (c), implying some information leakage.

1. **Hyper-parameters tuning**. Hyper-parameters are those parameters of the learning model that cannot be adaptively learned and must be set "by hand": there is no closed-form expression as for the regression coefficients, nor an iterative procedure. In the case of OQML, hyper-parameters are the σ's, representing the the Gaussian kernel width and regularization coefficient λ. In the case of MACE, there a several hyper-parameters settings how the neural network is optimised. In our study we varied the irreducible representations, radial cutoff, number of epochs and batch size, as well as the energy/forces weights ratios and at which point of the learning this ratio should change (swa). Details about the hyper-parameters can be found in Appendix C. They can be tuned by a grid search over different couples of values. The grid is exhaustively explored at low training subset size, then it is reduced to the most significant regions based on the best parameters found in the previous runs. For example, in the case of OQML the couple of hyper-parameters to optimize are $\boldsymbol{\xi}_k = (\lambda_i, \sigma_j)$.

2. **Model cross-validation**. For each point $\boldsymbol{\xi}_k$ of the grid parameters we run k-fold Cross-validation (CV). This means that the training subset is further divided into equally-sized and non-overlapping k sets, called folds, respectively (here $k = 3$), and the model is trained (KRR) on all of them, considered as a single train subset, except one (called validation set), which is used to test the performance (i.e. measure the error) and validate the model, that is, the hyper-parameters. The procedure is repeated k times, each one excluding a different fold from the training. The purpose of this is to compute the error on different folds and then considering the average.



3. The error definition depends on which labels we are training the model. In general it has the form

$$L_{\text{val}}(\boldsymbol{\xi}_k) = p\frac{1}{N_{\text{val}}}\sum_{i}^{N_{\text{val}}}(E_i^{\text{val}} - E_i^m(\boldsymbol{\xi}_i))^2 + (p-1)\frac{1}{3N_{\text{val}}M}\sum_{i}^{N_{\text{val}}}\sum_{a}^{M}\left\|\mathbf{f}_i^{\text{val}} - \mathbf{f}_i^m(\boldsymbol{\xi}_i)\right\|^2, \quad (7.18)$$

Where M represents the number of atoms, the superscript "val" refers to the true labels from the validation set (the subset momentarily excluded from training), while the superscript m indicates the labels predicted by the model. The coefficients p and $p-1$ are used to weight the contributions of energies and forces differently, creating a Pareto front. While this can be useful during the error evaluation in the fitting phase, we observed minimal performance differences for various p values at this stage of learning. Therefore, we chose $p = 0.5$. If the training is based solely on energy or forces, only the first or second term is used to compute the validation error, respectively.

Once the hyper-parameters have been selected, the training is done again on the whole training subset, without any division in k-fold, so that it exploits all the data at disposal.

Test. The model trained on the subset with the chosen hyper-parameters is tested on the test set which was put aside in the first place. The predicted energies and forces are compared with the true ones in what we called fitting error, $\boldsymbol{\rho}_E$ and $\boldsymbol{\rho}_{\mathbf{f}}$, in Section 7.1, using either root mean square errors (RMSE) or mean absolute error (MAE).

7.6 Results

7.6.1 Learning curves

The goal of the *learning curve* is to show the scaling of the performance of a model, measured as fitting error ρ, with fixed hyper-parameters with respect to the size of the training set, N_{train}.

$$\rho = \rho(N_{\text{train}}), \quad (7.19)$$

In this work for each training subset size we retrain the model from scratch, allowing the hyper-parameters to change and adapt depending on the configurations. This would measure how effective is the learning algorithm as a whole with respect to N_{train}.

An example of a learning curve is shown in Figure 7.9a, where we plot both the model errors, ϵ_E and $\epsilon_{\mathbf{f}}$, as well as the fitting errors, ρ_E and $\rho_{\mathbf{f}}$, the latter being the only metric available in standard settings.

One of the most noticeable aspects in the energy learning curves is the difference between the OQML and MACE trends. OQML exhibits a steady behavior from small training set sizes up to 1600 configurations, although training beyond this point with kernel methods becomes computationally demanding.

In contrast, MACE proved more difficult to optimize in the presence of noise for small training set sizes. As a result, the curve starts at 200 configurations, with both high model errors (ϵ_E) and test errors (ρ_E). However, once the model converges, its performance slightly surpasses that of OQML. Due to its implementation and design, MACE can be trained on larger datasets more easily than OQML, as shown by its learning curve extending up to 3200 configurations. The force learning curves show comparable behavior. Another prominent observation from these graphs is the significant gap between the fitting errors and model errors for both energy and forces, with ρ and ϵ following almost parallel trajectories.

In this study we are also interested in the *noise sensitivity curve*, that shows the influence of the underlying noise on the machine learning algorithm for fixed training set size, as in Figure 7.9b. The most striking property emerging from this plot is that the test error ρ is a rather pessimistic estimate of how far the model is from the ground truth, showed instead by the ϵ curves. This large difference could be a signal of the fact that we are in presence of good MLIPs models, that is, models that are only slightly affected by the QMC noise, as they are defined in [348]. However, we notice that for large input noise, the assumption of a linear relation between $\boldsymbol{\epsilon}$ and $\boldsymbol{\sigma}$,

$$\boldsymbol{\epsilon} = \eta\boldsymbol{\sigma} + \boldsymbol{\epsilon}_0, \tag{7.20}$$

does not hold anymore in the case of MACE energy prediction for $(\sigma_E, \sigma_f) > (77\,\text{meV}, 274\,\text{meV/Å})$, which is the point at which MACE performance on E becomes worse than the OQML one. Notice that crossover already happens in the "linear regime" in the forces plot, precisely at $(\sigma_E, \sigma_f) > (41\,\text{meV}, 145\,\text{meV/Å})$, which is an interesting point, since it is the noise level used in our QMC-driven molecular dynamics simulation, as reported in Table 7.1.

The observations made so far can be summarised using *learning tables* (Fig. 7.10), where both the influence of the noise and of the training set size on the fitting and model errors are plotted in a compact 2-dimensional plot. Indeed, the learning curves are the result of fixed-noise performances (rows), while the noise sensitivity curves are the result of fixed-training set-size performances (columns).

As noted in the curves described above, OQML shows a more steady performance with respect to N_{train} and noise levels than MACE (the lower scores on forces for $N_{\text{train}} = 800$ are probably due to sub-optimal exploration of the hyper-parameter grid in the validation phase). This result will be contrasted in the next Section, when we will measure performance on the base of physically sounded test, rather than based on statistics only. Again, MACE energies converge quite fast for level of noise up to $\sigma_E = 41\,\text{meV}$, with a number of configuration of $N_{\text{train}} = 400$ enough to reach the accuracy of noiseless-trained models (first row). Above that level of noise, MACE learning curves start to be less consistent, indicating a departure from the "linear regime" of Eq. (7.20).

The learning tables also potentially allow for a "diagonal" reading of the relationship between training set size and noise. As mentioned in Section 7.1, the level of precision σ^{V} of

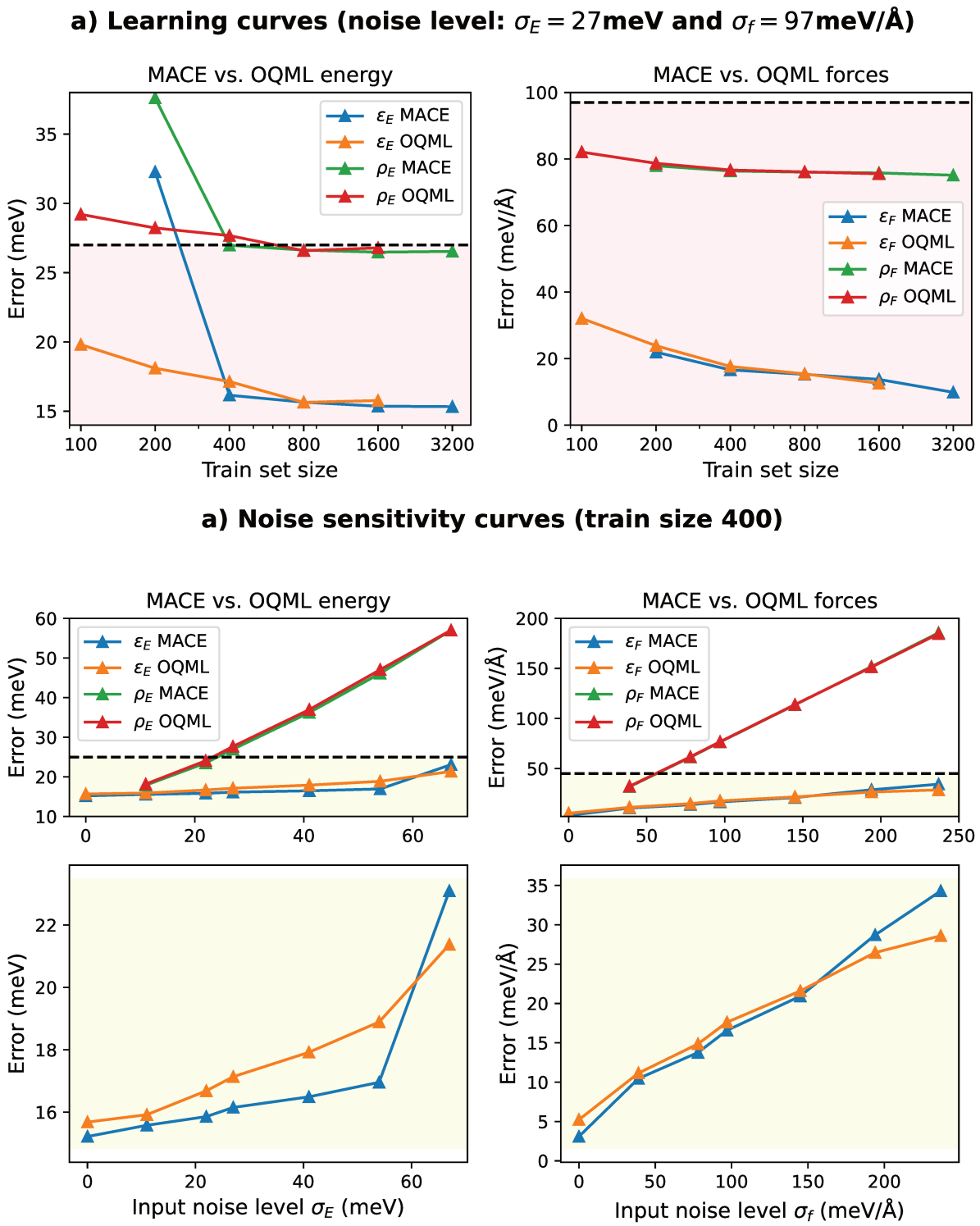


Figure 7.9: **(a) Learning curve for** $\sigma_E = 27$ meV **and (b) noise sensitivity curve for** $N_{\text{train}} = 400$**.** Results on the energies are reported on the left column, while the performance on forces are plotted in the right one. The second row of plots in the noise sensitivity (b) is an inset of what is shown in the graphs above, corresponding to a zoom on the same yellow area.

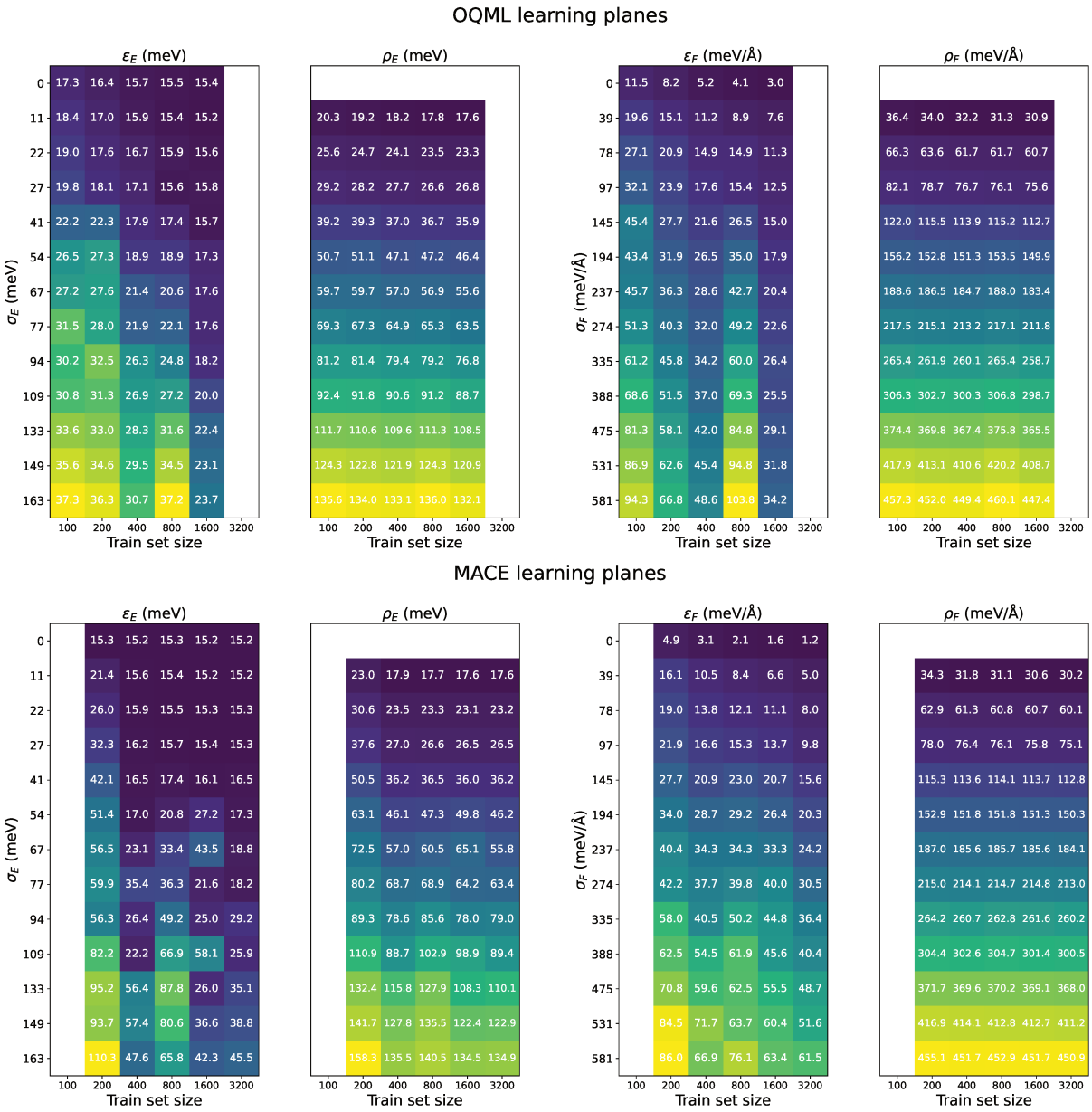


Figure 7.10: **Learning planes for (a) OQML and (b) MACE interatomic potentials.**

energies and forces QMC estimates depends on the number of electronic samples according to

$$\sigma^{\mathrm{V}} \propto 1/\sqrt{N_{\mathrm{gen}}}, \tag{7.21}$$

meaning that training sets characterized by the same $\sqrt{N_{\mathrm{train}}}/\sigma^{\mathrm{V}}$ ratio have equivalent computational costs. This comparison can help reveal whether precision or the exploration of the configuration space is more crucial for the model's performance. Improvements for larger training sets with lower accuracy suggest that exploring the configuration space is more important, while improvements for smaller training sets with higher accuracy indicate that the configuration space is already well-explored, and enhancing the quality of estimation should be the priority.

In our case, the fact that the energy error ϵ_E for the model trained on a dataset with noise below $\sigma_E = 54\,\mathrm{meV}$ rapidly converges to the energy error of the model trained on the clean dataset ([15.2-17.3] meV) limits the scope of this analysis. It is clear that further improvements must come from more advanced training set selection techniques, such as those mentioned in Fig. 7.8).

Nevertheless, this type of analysis can still be applied to the force predictions of the OQML model, as shown in Table 7.3, Here, we observe a slight improvement in performance, suggesting

Training set size	100	200	400	1600
σ_f(meV/Å)	97	145	194	388
ϵ_f(meV/Å)	32.1	27.7	26.5	25.5

Table 7.3: **Diagonal of ϵ_f in OQML.**

that better training set selection strategies would benefit not only energy predictions but also force predictions (we excluded the diagonal point (274 meV/Å, 800) due to a non-optimal grid search).

In the case of MACE, we quickly enter a non-linear performance regime, indicating that noise has a significant impact on the model's error. This is further supported by the observation that, although MACE's force errors are lower than those of OQML, its performance is approximately four times worse than MACE trained on a clean dataset (see the first two rows of ϵ_f in MACE learning table, Fig. 7.10):

$$\epsilon^{\mathrm{MACE}}_{f,\sigma=39\,\mathrm{meV/Å}} \approx 4\epsilon^{\mathrm{MACE}}_{f,\mathrm{clean}} \tag{7.22}$$

whereas for OQML, the performance degradation is only about twice as bad as the clean training:

$$\epsilon^{\mathrm{OQML}}_{f,\sigma=39\,\mathrm{meV/Å}} \approx 2\epsilon^{\mathrm{OQML}}_{f,\mathrm{clean}}, \tag{7.23}$$

(see the first two rows of ϵ_f in OQML learning table, Fig. 7.10).

7.6.2 Testing on physical observables

To relate how the noise in the learned PES impacts the MLIP-driven molecular dynamics we need a systematic way of comparing physical observables across their whole variability range.

Since most observables statistical distributions are computed as empirical histograms cumulated along the trajectory, we can test their compatibility following the method detailed in Ref. [369].

Let us describe the histogram of an observable collected during the dynamics as a collection of B couples of values, $\{(n_b, \sigma_b)\}_{b=1,\cdots,B}$ associated to each bin b, where n_b are the number of events in the bin and σ_b is an estimate of the standard deviation of that number, usually computed by block averaging techniques. Now consider two molecular dynamics simulation runs at the same conditions (for instance, the same temperature), with the only difference that one is based on the reference PES (α) and the other is based on the machine learned PES (β). Then we will have two histograms:

$$\begin{aligned}&(n_{1\alpha}, \sigma_{1\alpha}), (n_{2\alpha}, \sigma_{2\alpha}), \cdots (n_{B\alpha}, \sigma_{B\alpha})\\ &(n_{1\beta}, \sigma_{1\beta}), (n_{2\beta}, \sigma_{2\beta}), \cdots (n_{B\beta}, \sigma_{B\beta}).\end{aligned} \tag{7.24}$$

The *normalised significance of the difference* of two bins is defined as

$$S_b = \frac{n_{b\alpha} - K n_{b\beta}}{\sqrt{\sigma_{b\alpha}^2 + K^2 \sigma_{b\beta}^2}}, \tag{7.25}$$

where K is a normalisation factor, usually the ratio between the total volume of observations in the two histograms. In the denominator of Eq. (7.25), $\sqrt{\sigma_{b\alpha}^2 + K^2\sigma_{b\beta}^2} = \sigma$, where σ is the stochastic error of $n_{b\alpha} - K n_{b\beta}$, assuming that $n_{b\alpha}$ $K n_{b\beta}$ are two independent measures. Since we will deal mostly with already normalised histograms, which are obtained from MD simulations having the same number of time steps of the same duration, in most of our use of the formula above we will have $K = 1$. Then we can define a 2-dimensional measure of the distance, or similarity, between the two histograms as the average of the significance $\bar{S}$ computed on all the bins, and its variance:

$$\left(\bar{S}, \mathrm{var}\left[\bar{S}\right]\right) = \left(\frac{1}{B}\sum_b^B S_B, \frac{1}{B(B-1)}\sum_b^B (S_b - \bar{S})^2\right). \tag{7.26}$$

It is important to notice that the average should be computed only for those bins where at least one of the two histograms have a signal, otherwise we would underestimate $\bar{S}$. If the average significance of the difference $\bar{S}$ is lower than 3, meaning that the difference between observed quantities is within 3σ, we can say that the quantities are compatible. As it is good practice in molecular dynamics simulations, we produced multiple independent runs (4 in our specific case) for each temperature and type of PES. Then, we mediated $\bar{S}$ over the independent MD runs, at fixed temperature and PES. This further consideration provides a more reliable measure of the spread of the significance, because it is based on multiple physical tests, rather than just being based on the statistics of a single run of the ML-PES. Indeed, by running mutiple MD simulations with independent initial conditions we can explore a richer variety of configurations upon which the binned observables are computed.

The two observables that we considered are the pair correlation function, which give us a global picture of the compatibility of the simulations, and the 3-body correlation function between the two oxygens and the central proton, which is an important quantity for the study of the proton transfer, as we have seen in Chapter 4. These observables are detailed in what follows.

Pair correlation function

The Pair Correlation Functions (PCFs) are the distribution of the interatomic distances between couple of atoms belonging to specific elements (possibly the same element):

$$g^{(2)}_{\mathrm{XY}}(r) = \sum_{i=1}^{N_\mathrm{X}} \sum_{j=1}^{N_\mathrm{Y}} \left\langle \delta(|\mathbf{q}_i - \mathbf{q}_j| - r) \right\rangle \tag{7.27}$$

where $\langle\cdot\rangle$ is the ensemble average, practically computed along the single MD trajectory, δ is the Dirac delta function, and N_X and N_Y are the numbers of atoms of types X and Y, respectively. Usually, when dealing with liquids, this quantity is multiplied by the prefactor $1/N_\mathrm{X}N_\mathrm{Y}$, and divided by the volume of the spherical shell within $[r, r + \mathrm{d}r]$, yielding the radial distribution functions (RDF). Since we are more interested in comparing the distribution of the peaks of the PCFs between different simulations, dividing by the shell volume would only reduce the height of the peaks, making it more diffucult to visualize the differences. Therefore, we did not consider any prefactor. In Figure 7.11 we show an example of such histograms.

3-body correlation function of $\mathbf{O_1}$, $\mathbf{O_2}$ and $\mathbf{H^+}$

In Figure 7.12 we show the reduced coordinates used in the three-body correlation function, $g^{(3)}(\mathrm{O_1O_2H^+})$, which involves two oxygens, $\mathrm{O_1}$ and $\mathrm{O_2}$, and the central proton, $\mathrm{H^+}$. More precisely, $g^{(3)}$ measures the correlation between the distance of the two oxygens ($\mathrm{d_{O_1O_2}}$), and the relative position of the central proton with respect to the flanking oxygens atoms, as represented in Figure 7.12. Formally, we define it as

$$g^{(3)}_{\mathrm{O_1O_2H^+}}(x, y) = \left\langle \delta(\hat{d}_{\mathrm{O_1O_2}} - x)\delta(\hat{d}_{\mathrm{O_1H^+}} - y) + \delta(\hat{d}_{\mathrm{O_1O_2}} - x)\delta(\hat{d}_{\mathrm{O_2H^+}} - y). \right\rangle \tag{7.28}$$

As we saw in Chapter 4, this quantity is relevant in the study of the proton shuttling between two water molecules, and, given the reduced dimension of the Zundel cation, it exhaustively describes the PT mechanism in this system, as the are no solvation effects.

A graphical example of these correlation functions can be found in Figure 7.13, where we plot the 2D histograms using a density color code for the height of the bins. At variance with the 1D histogram of $g^{(2)}$, where the standard deviation is explicitly plotted, in this case we dedicated (NO: the rightmost column) two columns to the standardized difference between the values of the bins in MLIPs simulations with respect to the MBE-based ones. The difference is standardized as it has been described in Section 7.6.2, and the values up to 3σ are showed. Differences that are equal or larger than 4σ are all represented with the same color (violet), as we do not consider them statistically compatible.

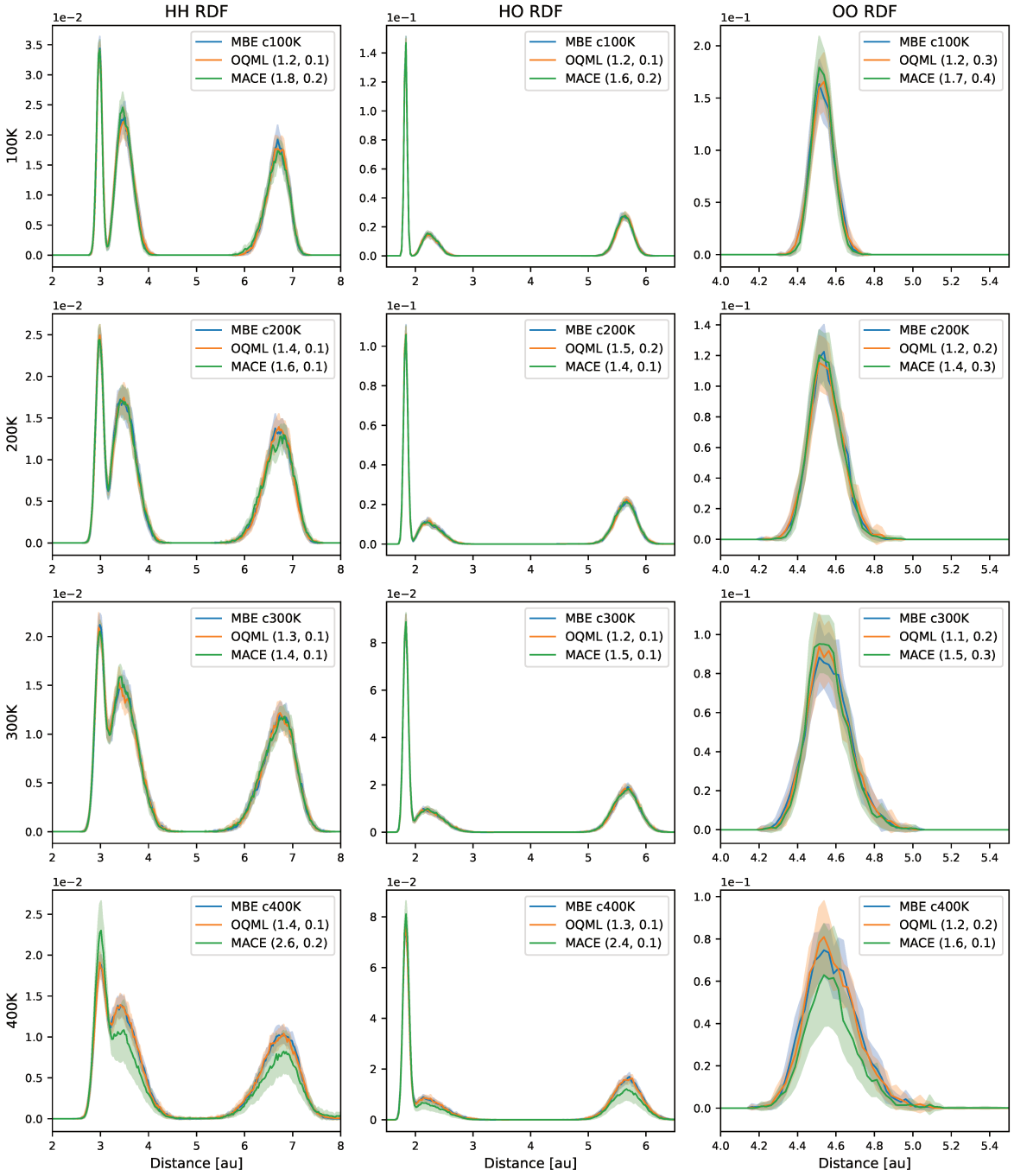


Figure 7.11: **Pair correlation functions of HH, HO and OO in the Zundel ion at different temperatures (classical simulations).**

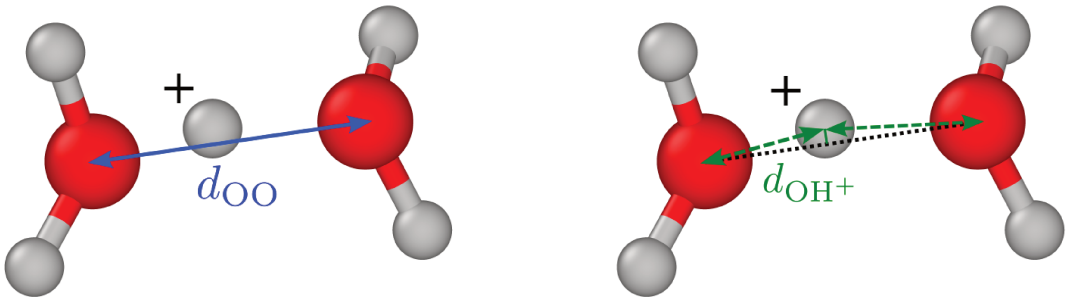


Figure 7.12: **Reduced coordinates for the study of the proton transfer in the Zundel ion.**

Normalised significance of the difference

In Figure 7.14 we plot the normalised significance of the difference in the case of the $g^{(3)}$. Analogous plots for other quantities conveys similar observations, therefore will be put them in Appendix C.

Independently from the noise level added to the training energies and forces, from a global point of view the first trends that we notice is the strong influence of the physical setup of the simulation, namely the temperature and the use of quantum nuclei, on the MLIPs performance. The training dataset was extracted from 300 K classical simulations, which makes the MLIPs reliable at lower or equal temperatures values, while at 400 K (d, h) most of the simulations are unstable. This is expected, as higher temperatures allow for a broader exploration of the configuration space, making more likely to run into a configuration which is far from the training dataset. At the same time, training on classical simulation translates into poorer performance in RPMD simulations, which show instabilities already in dynamics at the training temperature (Fig. 7.14g). These observations are generally valid also for MLIPs trained on deterministic datasets, and are expected already from the inspection of the dataset broadening in the dimensionality reduction shown in Figures 7.3d and 7.3e.

From the point of view of noise, in classical simulations at low temperatures (100 K and 200 K) both OQML and MACE show analogous robustness at any noise level. Differently, in quantum simulations at low temperatures their behaviour is quite different: MACE shows reliable results up to $(\sigma_E, \sigma_f) = (77\,\text{meV}, 274\,\text{meV/Å})$, with a ladder-shaped histogram profile showing decreasing performance, coherently with the increase of the noise level. On the other hand, while OQML performs very well with most of its normed difference bars way below the limit of 3σ, its behaviour is less predictable, as the shown by the 'spike' corresponding to the the simulation where the training dataset has been corrupted with noise level of $(\sigma_E, \sigma_f) = (54\,\text{meV}, 194\,\text{meV/Å})$ (Fig. 7.14f).

This unpredictability is found also in simulations at higher temperatures, like the one of the training set (300K, Fig. 7.14c,g), and in an extrapolating regime (400K, Fig. 7.14d,h).

or already at $\sigma_E = 11\,\text{meV}$ at the training temperature ((Fig. 7.14g). On the other hand, MACE looks more stable and predictable, showing monotonic decreasing performances as σ_E increases, remaining below the statistical significance line for $\sigma_E \leq 67\,\text{meV}$ at $T \leq 200\,\text{K}$, and for $\sigma_E \leq 22\,\text{meV}$ at $T = 300\,\text{K}$,

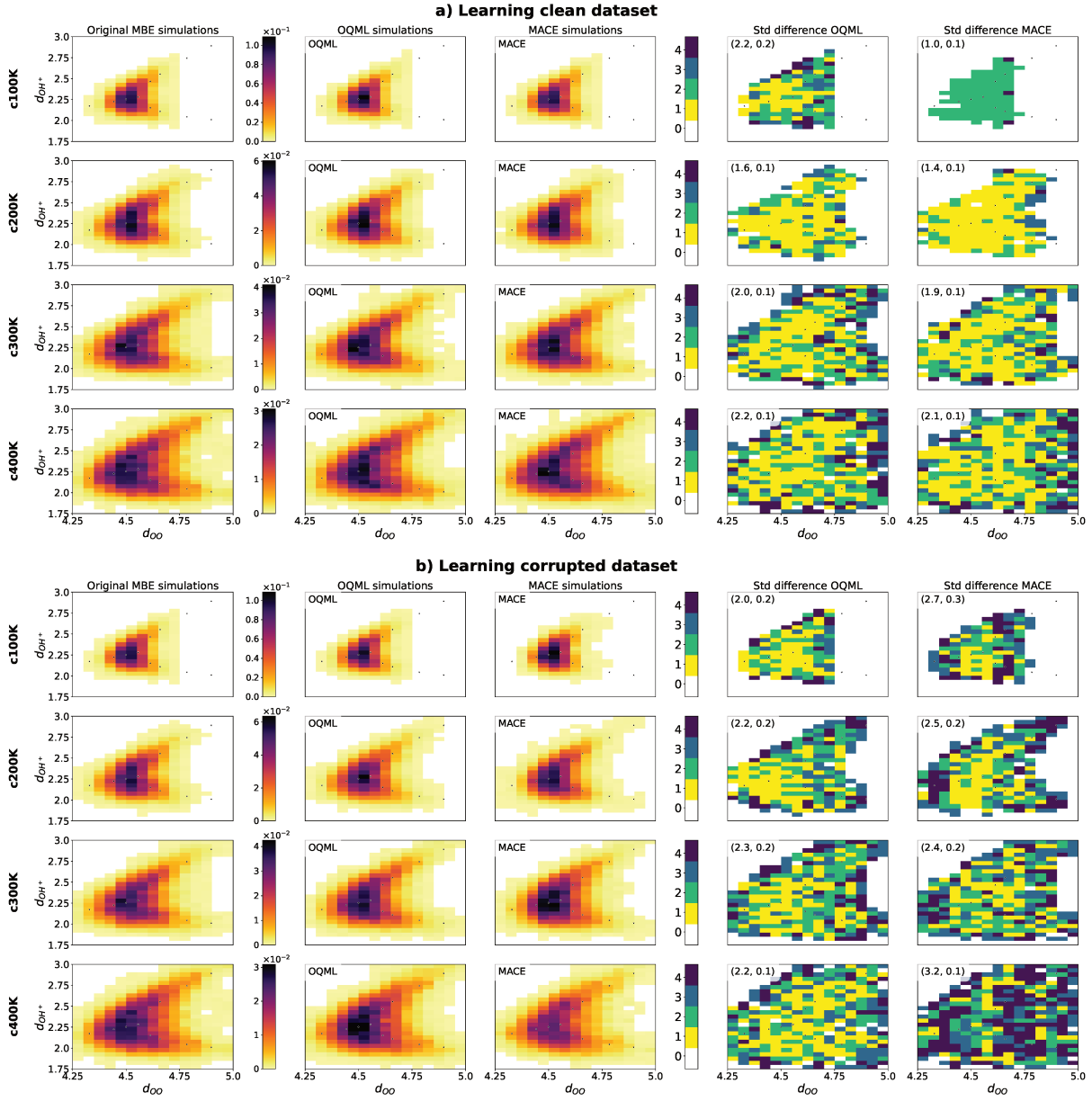


Figure 7.13: **Normalised 3-body correlation function 2d histograms obtained from MLIP-driven classical simulations.** The temperature varies for different rows, including [100 K, 200 K, 300 K, 400 K]. The first column represents the reference, that is, the results of the original MBE simulations. The second and third column are the results obtained with OQML-driven and MACE-driven dynamics, respectively. The density colormap is shared among plots on the same row to facilitate comparison, but it changes for different temperatures as the configuration distribution is T-dependent. The last two columns show the standardised difference $\overline{S}$ of the 2D histograms between OQML-driven vs MBE-driven dynamics, and MACE-driven vs MBE-driven dynamics, respectively. For these plots the colorcode reflects the σ_S. Values above 3σ are all considered statistically incompatible and plotted with the same colour (violet). In a) the results are obtained for a MLIPs trained on the original, clean MBE dataset, not affected by additional artificial noise. In b) the MLIPs were trained on a dataset corrupted with a $\sigma_E = 11$ meV level of noise. The difference

Although the MACE normalised significance of the difference is compatible for most of the noise levels in classical simulation at 100 K and 200 K, it is interesting that classical simulations show slightly worse performance at 100 K than 200 K, with highest differences for $\sigma_E = 11\,\text{meV}, 22\,\text{meV}, 109\,\text{meV}$, and the quantum one as well for $\sigma_E \leq 77\,\text{meV}$. While in the latter case this can be explained by the wider extension of the necklace at the lowest temperature (128 beads against 64), as it was shown also in Figure 7.3d, in the classical case could be a symptom of the fact that despite being in an interpolation regime (the classical configuration space explored at 100 K is a subset of the one explored at 300 K), the simulation could be more sensible to the precision with which the PES shape is reproduced. Again, the solution for this could be mixing datasets produced at different temperatures, especially exploiting the fact that at low temperatures the QMC wavefunction updates from on step to another of the dynamics are less drastic than those at higher temperatures.

7.7 Preliminary results on the protonated water hexamer

Preliminary work on applying MLIPs to the protonated water hexamer was conducted as a validation of the results obtained from QMC-PILD simulations (Chapter 4). Although our QMC-trained MLIP for $H_{13}O_6^+$ did not yet reach the robustness needed for long simulations, it proved useful in identifying a subsampling issue in the original QMC-MD. Specifically, the classical QMC-driven simulation at 100 K displayed a peculiar symmetrized radial distribution function of the central proton relative to the two oxygens. Further inspection, through comparison with the OQML-driven dynamics, confirmed that this was an artifact.

The initial attempts to study the effectiveness of learning $H_{13}O_6^+$ followed a strategy similar to that used for the Zundel ion: comparing the performance of an MLIP trained on a noiseless, deterministic dataset against the same MLIP trained on QMC data.

This type of comparison can be done by analysing learning curves, but it does not allow for a detailed investigation of the noise's effect on the physics of the model, as the underlying PES sampled by the two electronic structure methods is inherently different. In fact, when comparing physical observables accumulated over MD trajectories based on different PES, it becomes challenging to distinguish the effects due to the influence of noise from the effects of a different level of theory. The analysis is further complicated by the presence of multiple isomers, with transition rates between them varying across different PES. This adds another layer of difficulty when comparing observables.

Nevertheless, we present here some of the key results observed.

Datasets

Starting from the same $H_{13}O_6^+$ configurations generated by QMC-driven MD simulations (Chapter 4), we calculated the energy and forces at the DFT level using `Quantum ESPRESSO` (QE) [370, 371]. The DFT calculations have been carried out using the PBE functional, corrected for the inclusion of van der Waals interactions [89], as mentioned in Chapter 2.

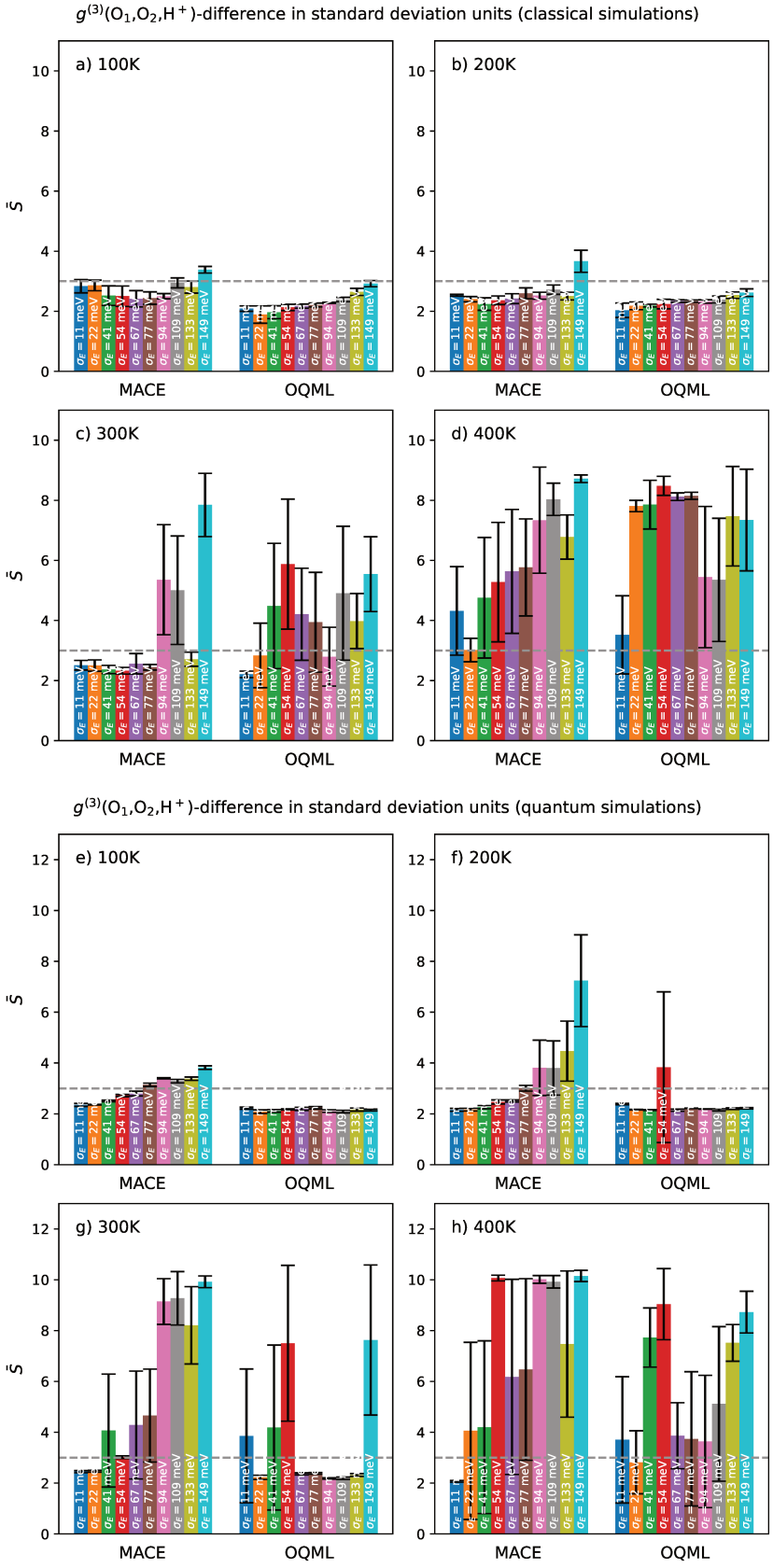


Figure 7.14: **Significance of the difference of $g^{(3)}$ between MD simulations based on MBE and MLIPs.** Blocks a,b,c,d refer to classical simulations, while blocks e,f,g,h refer to quantum (ring polymer) simulations.

Within this computational framework, two issues arise: firstly, we are studying an isolated cluster using periodic boundary conditions imposed by the periodic basis set used by QE. Secondly, we are dealing with a charged cluster, which, despite the relatively simple monopole-like term of the hydrogen ion, still implies the presence of long-range interactions.

For such an isolated charged molecule, a large simulation cell is needed to model the surrounding vacuum and to avoid interactions between periodic replicas. Moreover, as the system is non-periodic, a high energy cutoff for the plane-wave expansion must be employed.

To address the challenge of simulating a charged system under periodic boundary conditions, two solutions have been proposed: the Makov-Payne (M-P) correction [372], which corrects the total energy, and the Martyna-Tuckerman (M-T) correction [373], which corrects both the total energy and the self-consistent potential. In this study, we applied the Martyna-Tuckerman correction as it offered greater stability, with the error either decreasing or remaining at the same order of magnitude between consecutive runs.

MLIP: kernel ridge regression methods

We used the FCHL19 representation in a kernel ridge regression framework, using several types of kernel, all described in Section 6.4: from energy-only learning (Sec. 6.4.2), to Gaussian process-type kernels for energies and forces (Sec. 6.4.3), to operator quantum machine learning (Sec. 6.4.4).

The resulting learning curves are reported in Figure 7.15. We observe that the DFT-OQML learning curves exhibit the expected power-law behavior described by Eq. (7.8), in contrast to the QMC-OQML learning curves, which appear flatter on a log-log scale. From the results on the Zundel complex (Section 7.6.1), we know that flat learning curves in the fitting error (ρ) may still hide some effective model training, at least before overfitting. This is partially evidenced by the fact that, despite the flatness of ρ_f, the test error on energies (ρ_E) continues to improve.

In both DFT and QMC learning, including forces during training proves invaluable, as it significantly enhances energy prediction accuracy and accelerates error convergence. Additionally, models trained solely on forces (yellow curves) can accurately reproduce energies through simple integration, while deriving the forces from energy only training is a much more diffucult task.

The best-performing OQML potential, according to the test error metric, was a model trained on 720 configurations selected via farthest point sampling applied to the 250 K dataset. We used this model in real dynamics simulations and computed the resulting radial distribution functions in both classical simulations (Figure 7.16a) and ring-polymer molecular dynamics (Figure 7.16b).

Interestingly, as with the Zundel complex, the PCFs at low temperatures (50 K)—both classical and quantum—are among the worst reproduced. This suggests that even in regions of the PES reachable by simple thermal fluctuations, the dataset may have gaps, an issue that can be exacerbated in RPMD simulations.

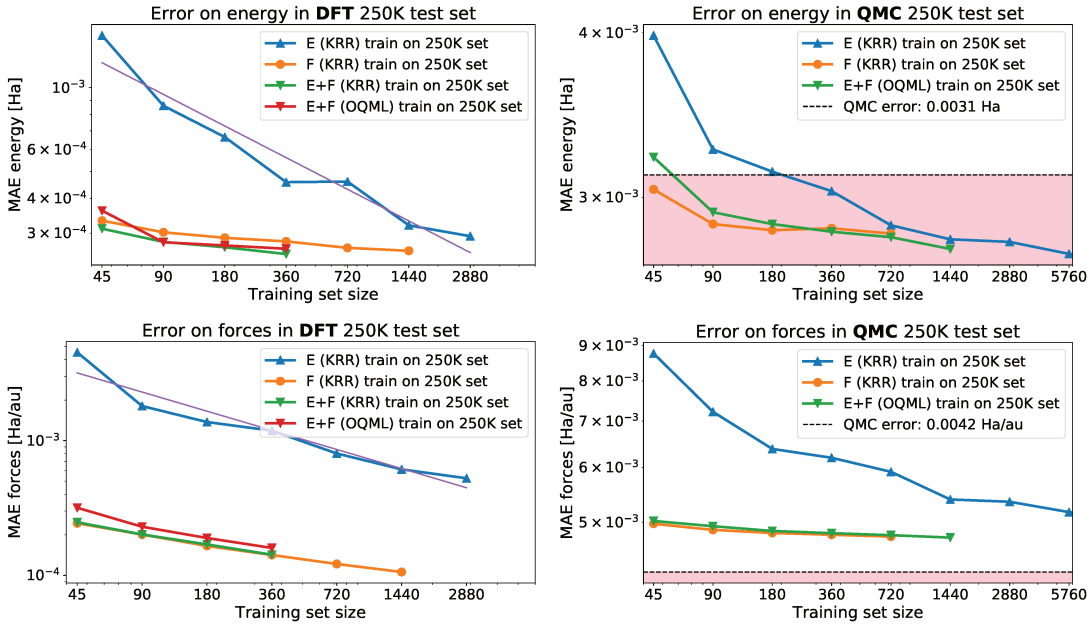
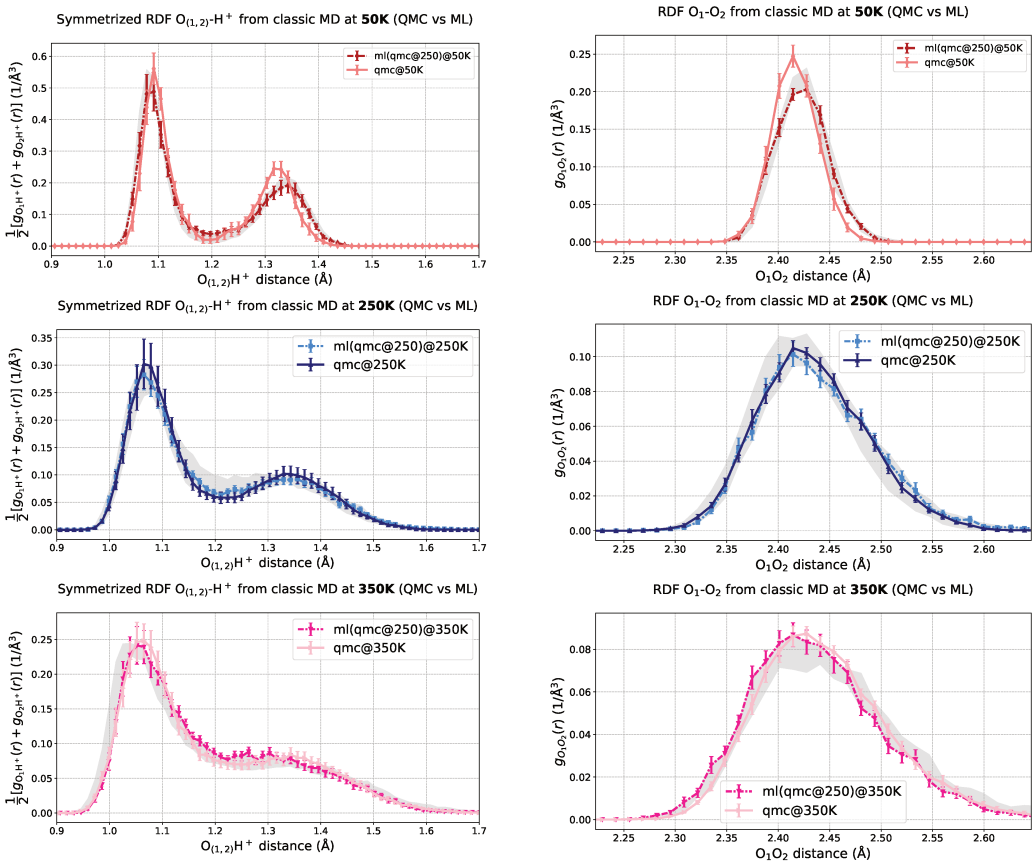
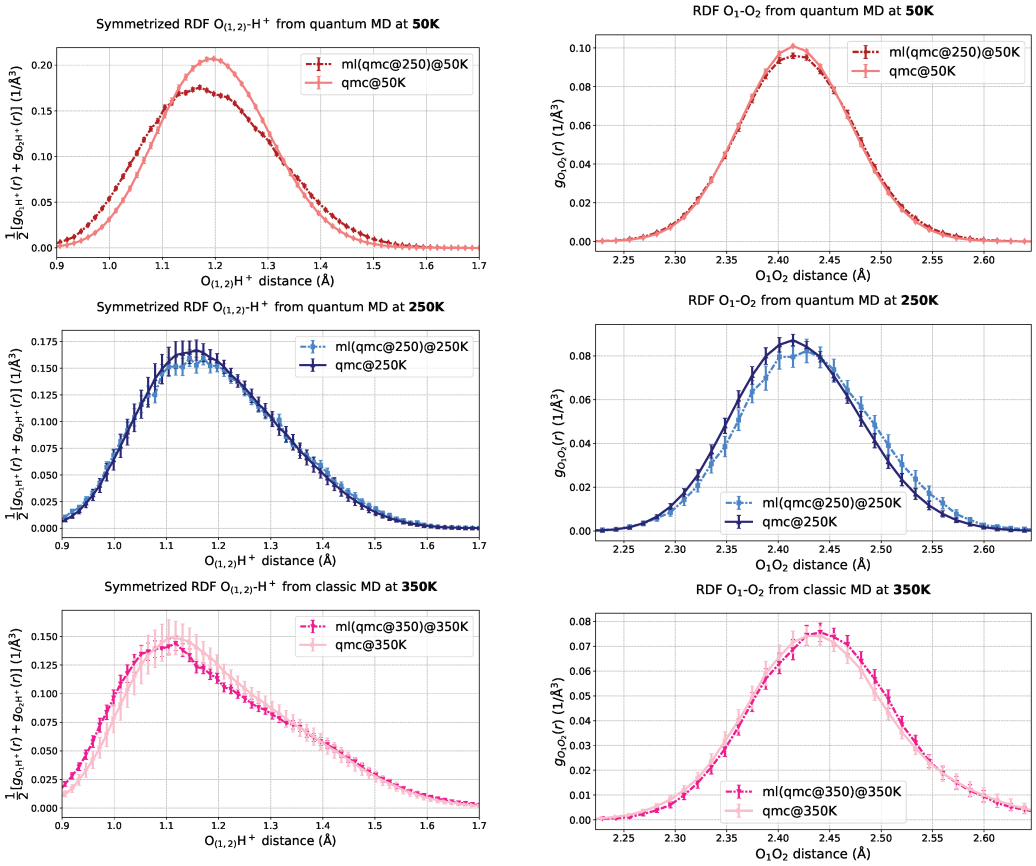


Figure 7.15: **Learning curves of KRR schemes trained on $H_{13}O_6^+$ configurations treated at DFT level (left), and QMC level (right).** Top plots refers to error in energy prediction, while bottom ones to the forces prediction. Blue curves are trained on energy only, yellow ones on forces only. E+F refers to energy and forces learning, which is done either using a Gaussian process regression-type of kernel (green curves in the DFT plots on the left), or relying on operator quantum machine learning (in red in the DFT plots on the left, in green in the QMC plots on the right). As we explained in the previous Chapter, OQML is less computational intensive than GPR-kernel based KRR, and since they show the same errors, we decided to focus on OQML.

Another known issue arises in the extrapolation regime, where the dynamics enters configurations too far from those seen during the training, leading to a breakdown of the simulations. In many cases, this occurred due to an incomplete description of the solvation shell, where water molecules are more flexible and prone to exploring unseen regions of the configuration space. The OQML-driven trajectory often brought to the formation of a Eigen-like hexamer, $H_9O_4^+(H_2O)_2$, that is, an unseen but plausible isomer. However, study of the trajectory revealed that the isomerisation path was not the correct one, often ending with losing one water molecule. Including different isomers in the training set could help the potential in dealing with edge configurations and improve the overall dynamics in the region of main interest. For example if we are more interested in accumulating instanton statistics from the Zundel configuration in a long simulation, it may help the potential know how to behave at the boundary of the interpolation region.



(a) **PCF in $H_{13}O_6^+$ in classical QMC-driven and OQML-driven simulations.**



(b) **PCF in $H_{13}O_6^+$ in quantum QMC-driven and OQML-driven simulations.**

Figure 7.16: **Radial distribution function in OQML-driven MD simulations of $H_{13}O_6^+$.**

Chapter 8

Conclusions

The purpose of this thesis was twofold: on the one hand we aimed at studying Proton Transfer (PT) in the protonated water hexamer by means of advanced computational methods, namely quantum Monte Carlo (QMC) and path integral molecular dynamics (PIMD). On the other hand, we assessed the robustness of current machine learning interatomic potentials (MLIPs) in fitting accurate but noisy potential energy surfaces (PES) estimated through QMC and PIMD, with the final goal of producing a QMC-trained PES for protonated water clusters. These goals are two sides of the same coin, the one of accuracy versus computational cost trade-off.

In the first part we showed the delicate interplay between thermal and NQEs in determining the range of temperatures at which proton transfer is optimal. We found that we have to include in the picture the contribution coming from the short-Zundel configurations, which are enhanced in quantum simulations. This result can be cast in a long-standing debate around the identity of the hydrated proton, and in this context we proved the need of going beyond the simplicistic classification into Zundel and Eigen moieties. Recent computational [374] and experimental [194, 195] findings also go in this direction.

Research on water systems has always exploited cutting-edge computational methods. This means that further progress in understanding proton solvation and diffusion in water will require accurate and unbiased PES as those delivered by QMC, together with the inclusion of NQEs. To overcome the computational burden of these methods, MLIPs are the most promising path. Motivated by the necessity of extending our simulations both in system size and simulation time, we investigated resilience of MLIPs, designed for noiseless data, when trained on energies and forces affected by a known level of noise, as in the QMC case.

Using the Zundel cation $H_5O_2^+$ as a benchmark, we built the "reference" training set by sampling configurations through MBE potential-driven. Then, we corrupted the dataset by adding errors sampled from multivariate Gaussian white noise characterised by increasing levels of variance (σ^2). We chose two types of ML potential, one based on kernel ridge regression, OQML, and one based on message passing neural network framework, MACE. The final goal was to check their reliability in production runs, and if it is possible to measure it in advance. To this end, we found out that the fitting error, referred to also as test error ρ, can be a poor metric,

as already pointed out in Refs. [350, 351], and this seems particularly true when learning QMC. Indeed we proved that the distance between the MLIP predicted energies and forces and the true ones (model error ϵ) is lower than the test error measured between the predictions and the noisy labels. The naive test error is useful as a first indication on the success of the learning, and can signal if more configurations should be included in the training set.

To go beyond learning curves analysis, we established a rigorous and standardised way of evaluating the quality of the simulations based on physical observables such as radial distribution functions or higher order correlation functions ($g^{(2)}$ and $g^{(3)}$, respectively, from Chapter 7) and which are particularly important in the system at hand. Using this metric, we found out that, at variance with OQML, MACE shows a more predictable behaviour with respect to the input noise affecting the training set. By predictable we mean that the quality of the MACE-driven dynamics degradates almost monotonically at increasing levels of noise, which is an appreciable property. OQML, despite showing even better performances at the low temperatures regimes, can "break" the simulation for unexpected levels of input noise.

Perspectives

Thanks to years of research on variance reduction and to the application of automatic differentiation, estimating accurate forces with QMC comes with a slightly higher cost than the one required by the energies. Also interpolating the PES and its gradient translates into higher computational needs, but the usefulness of derivative information in learning such complex and high-dimensional surfaces, in contrast to methods based on energy only, is out of question. This is apparent in the learning curves of the protonated water hexamer, where the inclusion of the forces makes the convergence on the energy error much faster than the models trained only on energies. The vectorial information contained in the forces is not merely quantitative, but it is qualitatively different from, for instance, simply including more energy points. Indeed, far from being just additional data with respect to plain, scalar energies, gradients provide a "smoother" view on the energy landscape. This is even more important when learning only noisy energies, as applying the derivative operator on the interpolated surface would potentially increase the errors [375, 376].

Moreover, within a stochastic electronic structure method, including forces in the training set rather than energies can become more and more convenient as the system size increases. Indeed, while the energy scales extensively with the system size, and so its error, the force information stays local, and its relative error distribution follows the dynamical matrix eigenvalues, as we have quantitatively verified in this thesis.

A further direction for improvement could involve generating training sets with optimally distributed configurations. In this study, classical simulations were employed to reconstruct the PES, as they allow nuclei to move freely without the interference of zero-point energy (ZPE), which distorts the true shape of the PES. However, ring polymer molecular dynamics (RPMD)

also offer an intriguing pathway: exploiting the extension of the necklace towards regions that would be otherwise inaccessible or unlikely to be explored in classical simulation just by thermal fluctuations. Specifically, since NQEs enable protons to more frequently overcome barriers, RPMD simulations may provide richer statistical data to better capture the barrier shape, although ZPE effects must still be accounted for.

Since selecting the best configuration for PES learning is a common problem when applying MLIPs, much of the research effort in the ML community has been devoted to find recipes to construct or improving training set-. The most straighforward one, only partially explored in the preliminary work on the hexamer, is farthest point sampling (FPS), where the training set is sparsified by sampling the most different configuration in terms of descriptors, or by maximising the kernel distance in some high-dimensional feature vector space.

A promising alternative is *active learning* based on a measure of uncertainty of prediction [377–379]: the higher is the uncertainty in some region of the PES, the more configurations from this very same region should be included in the training set. In contrast with *a posteriori* active selection of the training configuration, another possibility is represented by an *on-the-fly* active sampling through uncertainty-driven dynamics [380], a method that has been dubbed "hyperactive learning" [381]. This is based on the idea that the Boltzmann weights that characterise MD-sampled distribution are not ideal when it comes to MLIP training. The appeal of these methods also lies in their ability to measure uncertainty [382], which would be interesting to compare with the noise of QMC-based datasets. This is especially important given the need to go beyond test error when evaluating a model's performance.

Recently, a simulation-oriented training has been proposed, which depends on (i) 'property-based metric' to describe the quality of the simulations, and on (ii) an optimization strategy based on the same metric [383]. This physically motivated construction of the training set relies on measuring how well an observable is reproduced with respect to the reference, implying that the observable itself must not be computationally intensive to evaluate. We believe that this *property-based training* points in the direction we took when we pushed the MLIP analysis toward molecular dynamics quality testing. In fact, once we know that $g^{(2)}$ and $g^{(3)}$ are well reproduced in MLIP-driven dynamics of $H_5O_2{}^+$ and $H_{13}O_6{}^+$, we can run longer simulations to gather more statistics about instantons in these systems.

Another important aspect to investigate when applying MLIPs to QMC data is how the intensity of noise impacts the learning algorithm in relation to the size of the system. For instance, with a fixed training set size, does a larger number of atoms with similar local environments lead to error compensation, thereby facilitating learning, or does the complexity of the PES outweigh the benefits of having more similar atomic environments?

Answering to the previous question would improve our understanding of the relation between noise and size of the system, especially in the perspective of learning a QMC-level of theory PES in order to improve the statistics of our QMC-driven simulations, building upon the findings that we reported in the first part of this thesis.

As stated in the Introduction, studying the proton transfer in bulk water is not just a matter of longer simulations based on very accurate cluster-PES. One should also take into account the solvation effects, hence considering the influence of adding more water molecules, read solvation shells [199], around the proton. Embeddings of such kind are usually constructed using quantum mechanics/molecular mechanics schemes (QM/MM) [384], which has been recently applied to the problem of PT in bulk water [385]. Given its nature of fluxional defects that propagates across the H-bond network of bulk water [48], the hydrated proton embedding is not an easy task, requiring on-the-fly adaptive partitioning [386, 387].

While there are limited studies on QMC/MMpol applications [388], some ML/MM framework have been recently proposed, like the deep potential range correction (DPRc) [389], successfully employed also in PIMD simulations [390], or electrostatic embedding of arbitrary MLIPs trained on molecular system *in vacuo* [391], which indeed is our case. Graph neural networks have been tested in ML/MM settings [392], as well as the ANI neural network potential [393].

Whether these solutions can be widely adopted will also depend on advances in the development of MLIPs capable of handling long-range interactions, particularly electrostatics, as discussed in Chapter 6 in our brief review of water potentials. This appealing approach would allow one to combine the accuracy and the speed of a QMC-MLIP core region with a molecular mechanics solvation environment.

Appendix

Appendix A

Stochastic integration schemes

In this Appendix, we provide the useful formulae to understand the algorithmic developments introduced in the Chapter 3.

A.1 Solution to the Ornstein-Uhlenbeck process for the Bussi algorithm

We aim at solving the Ornstein-Uhlenbeck process stochastic differential equation. Its differential form is the following:

$$p(t + \mathrm{d}t) - p(t) = \mathrm{d}p = -\gamma p(t)\,\mathrm{d}t + B\,\mathrm{d}W(t). \tag{A.1}$$

To simplify the notation, we use scalar quantities, and we ignore the possible dependences on the positions, like the one of $B = B(q)$.

To solve the equation, we consider the expectation values and the variance of the random variable p:

$$\mathbb{E}\left[p(t + \Delta t) - p(t)\right] = -\gamma\,\mathbb{E}\left[p(t)\right]\mathrm{d}t \quad \rightarrow \quad \mathbb{E}\left[p(t + \Delta t)\right] = p_0 e^{-\gamma \Delta t}, \tag{A.2}$$

$$\begin{aligned} \mathrm{var}\left[p(t)\right] &= \mathbb{E}\left[p(t + \Delta t)^2\right] - \mathbb{E}\left[p(t + \Delta t)\right]^2 \\ &= \mathbb{E}\left[p(t + \Delta t)^2\right] - p_0^2 e^{-2\gamma \Delta t}. \end{aligned} \tag{A.3}$$

In order to find the first term of the previous equation, we write

$$\begin{aligned} \mathrm{d}[p(t)^2] &= [p(t + \mathrm{d}t)]^2 - [p(t)]^2 \\ &= [p(t)(1 - \gamma\,\mathrm{d}t) + B(q(t))\,\mathrm{d}W]^2 - [p(t)]^2 \\ &= -2p(t)^2\gamma\,\mathrm{d}t + 2p(t)B(q(t))\,\mathrm{d}W + B^2(\mathrm{d}W)^2. \end{aligned} \tag{A.4}$$

Taking the expected value of the above quantity, we get

$$\mathrm{d}\,\mathbb{E}\left[p(t)^2\right] = -2\,\mathbb{E}\left[p(t)^2\right]\gamma\,\mathrm{d}t + B^2\,\mathrm{d}t, \tag{A.5}$$

where we used two properties: (1) the fact that $\mathrm{d}W(t)^2 = \mathrm{d}t$, and (2) the fact that $p(t)$ and $\mathrm{d}W(t)$ are statistically independent, therefore the expectation value of their product is the product of their expectation values, which is zero for $\mathrm{d}W$.

Differentiating the previous equation we get:

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{E}\left[p(t)^2\right] = -2\,\mathbb{E}\left[p(t)^2\right]\gamma\,\mathrm{d}t + B^2\,\mathrm{d}t \tag{A.6}$$

with solution

$$\mathbb{E}\left[p(t)^2\right] = p_0 e^{-2\gamma t} + \left(\frac{B^2}{2\gamma}\right)\left(1 - e^{-2\gamma t}\right) \tag{A.7}$$

Inserting what we have found in the initial equation for the variance, we finally obtain:

$$\mathrm{var}\left[p(t)\right] = \left(1 - e^{-2\gamma t}\right) \tag{A.8}$$

By combining the average and a random Gaussian vector multiplied by the above factor, we obtain the first step of the Bussi algorithm.

Appendix **B**

2D projection of the protonated hexamer PES

B.1 Towards an accurate modeling of the potential energy surface

We exploit the calculation of VMC forces not only to perform QMC-driven classical and quantum LD, but also to extract the best PES fitting functional form for the excess proton and for the water-water interaction in the Zundel core. The final goal is to derive the two-dimensional (2D) model potential $V_{2D} = V_{2D}(d_{O_1O_2}, \delta_{H^+})$, where $d_{O_1O_2}$ is the distance between the two central oxygen atoms and δ_{H^+} is the proton sharing coordinate, referenced to the midpoint of the $O_1H^+O_2$ complex: $\delta_{H^+} \equiv \tilde{d}_{O_{1/2}H^+} - d_{O_1O_2}/2$, with $\tilde{d}_{O_{1/2}H^+}$ the $\overline{O_{1/2}H^+}$ distance projected onto the $\overline{O_1O_2}$ direction. The projection of the full interatomic potential on the restricted 2D manifold is done by integrating the other degrees of freedom over the thermal partition function, sampled during the MD, i.e.

$$V_{2D}(d_{O_1O_2}, \delta_{H^+}) \equiv \left\langle V(x_1, x_2, \cdots, x_{3N}) \delta(x_1 - d_{O_1O_2}) \delta(x_2 - \delta_{H^+}) \right\rangle, \tag{B.1}$$

where $\langle ... \rangle$ is the average over the partition function of the classical/quantum statistical ensemble at fixed temperature, and V is the $3N$-dimensional potential depending on the generalised nuclear coordinates of the full system, $\mathbf{X} = (x_1, x_2, \cdots, x_{3N})$.

Analogously, one can define the one-dimensional (1D) potential acting between O_1 and O_2 as

$$V_{1D} = V_{1D}(d_{O_1O_2}) \equiv \left\langle V(x_1, \cdots, x_{3N-2}) \delta(x_1 - d_{O_1O_2}) \right\rangle, \tag{B.2}$$

according to previous notations. Derivatives of the previous potentials with respect to $d_{O_1O_2}$ and/or δ_{H^+} can be defined in the same way. For instance,

$$\frac{\partial V_{2D}}{\partial \delta_{H^+}} \equiv \left\langle \frac{\partial V(x_1, x_2, \cdots, x_{3N-2})}{\partial x_2} \, \delta(x_1 - d_{O_1O_2}) \delta(x_2 - \delta_{H^+}) \right\rangle, \tag{B.3}$$

and

$$\frac{\partial V_{1D}}{\partial d_{O_1O_2}} \equiv \left\langle \frac{\partial V(x_1, x_2, \cdots, x_{3N-2})}{\partial x_1} \, \delta(x_1 - d_{O_1O_2}) \right\rangle. \tag{B.4}$$

Given these definitions, we can proceed with the calculations of the corresponding quantities with the aim at modeling the potentials V_{1D} and V_{2D}. To do so, we will integrate the other degrees of freedom using the classical Boltzmann distribution in $\langle ... \rangle$, as generated by the QMC-driven classical Langevin dynamics at 100, 250 and 350 K. Employing the classical partition function has the advantage that the potentials *sampled* in this way will tend to the original PES of the system as $\beta \to \infty$, while the quantum partition function will lead to averaged potentials biased by quantum fluctuations even in the zero temperature limit. To compute these quantities from an MD sampling, the δ-functions in their definitions above are replaced by bins, whose size is given by the spacing between neighbouring points.

In Fig. B1, we study the $V_{1D}(d_{O_1O_2})$ potential depending on the water-water distance $d_{O_1O_2}$ (left column), and its derivative $\partial V_{1D}/\partial d_{O_1O_2}$ (right column). As one can see, the energy profile, at the left-hand side, is much more noisy than the behavior of its gradient, from where we can extract a precise value of the equilibrium $d_{O_1O_2}$ distance, and the evolution of the potential around the minimum. This shows the advantage of computing QMC forces in order to determine the PES, and suggests that a robust way of deriving the V_{1D} potential is by fitting and integrating its derivatives, rather than by directly fitting the energies, as pointed out also in machine learning potentials interpolating noisy PES (Chapter 7).

In Fig. B2, we study the $V_{2D}(d_{O_1O_2}, \delta_{H^+})$ potential depending on the proton coordinate δ_{H^+}, at various (fixed) $d_{O_1O_2}$ distances. In the left column, we show $\partial V_{2D}/\partial \delta_{H^+}$ in a contour plot as a function of both $d_{O_1O_2}$ and δ_{H^+}. Positive (negative) values of $\partial V_{2D}/\partial \delta_{H^+}$ are coloured in red (blue). The white region indicates the extrema of the 2D-PES. The classical proton is clearly asymmetric for $d_{O_1O_2} \gtrsim 2.37$ Å, with a minimum departing from the $\delta_{H^+} = 0$ axis. In the right column, the same information is provided by superposing $\partial V_{2D}/\partial \delta_{H^+}$ plotted as a function of δ_{H^+} and taken at fixed $d_{O_1O_2}$ distances.

B.2 Projected two-dimensional PES

Using the data obtained in Sec. B.1, let us determine an analytic form for the $V_{2D}(d_{O_1O_2}, \delta_{H^+})$ potential, which depends on both $d_{O_1O_2}$ and δ_{H^+} coordinates. This will take into account the variation of the proton-oxygen potential along the proton shuttling mode as the distance between the two inner water molecules varies.

We first derive the V_{1D} potential between the two water molecules, which depends only on the $d_{O_1O_2}$ stretching coordinate, by fitting the derivatives shown in Fig. B1, for the simulation at 100 K, which yields less noisy datapoints than the one at higher temperatures. As fitting function, we choose the Morse potential, such that:

$$V_{1D}(x) = D\left(1 - e^{-w(x-d_e)}\right)^2 \tag{B.5}$$

where we have chosen to set the zero of energy at the potential minimum, that is, at the equilibrium distance d_e; D and w represent the depth and the width of the potential well, respectively.

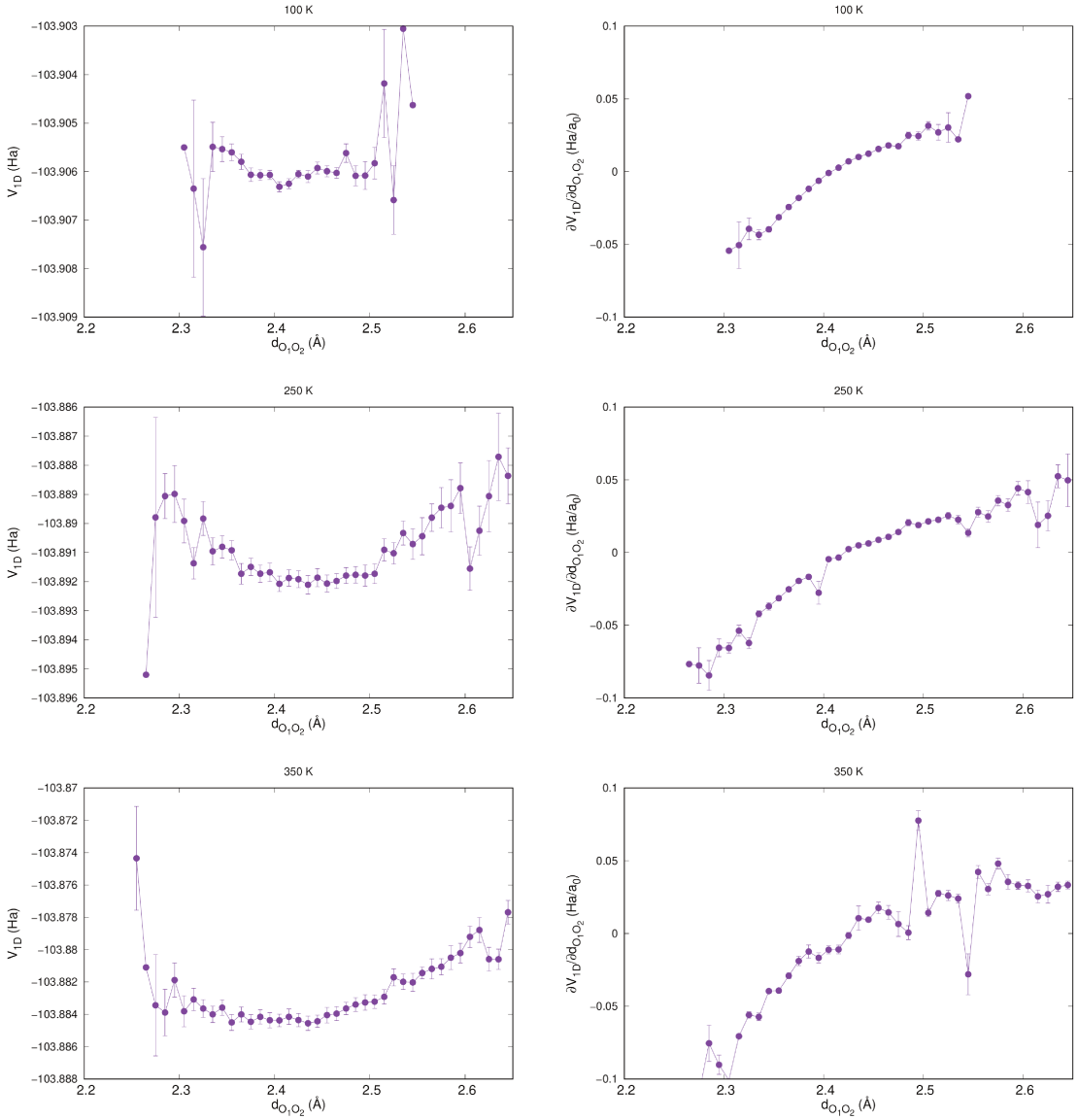


Figure B1: **Left-hand side: total energy variation of the cluster as a function of the $d_{O_1O_2}$ distance (V_{1D}). Right-hand side: its derivative, $\partial V_{1D}/\partial d_{O_1O_2}$, resulting in the force that drives the O_1-O_2 stretching mode.** The latter is computed as the sum of the energy gradients with respect to $\mathbf{q}_{O_1}$ and $\mathbf{q}_{O_2}$ variations projected along the O_1O_2 direction, for classical simulations at different temperatures.

The results of the fit are plotted in Fig. B3a, together with the potential derivatives evaluated by classical MD driven by QMC forces at 100 K, 250 K and 350 K.

From this analysis, the estimated equilibrium distance between two water molecules in the Zundel core of the protonated water hexamer is 2.408 Å at 100 K, in good agreement with the analysis based on the radial distribution function reported in Fig. 4.4 of the main text.

While the data derived from QMC-MD simulations are less noisy at 100 K, they however explore a smaller phase space, due to a probability density distribution more localised in the $(d_{O_1O_2}, \delta_{H^+})$ space at lower temperatures. This turns out to be a problem, if one aims at estimat-

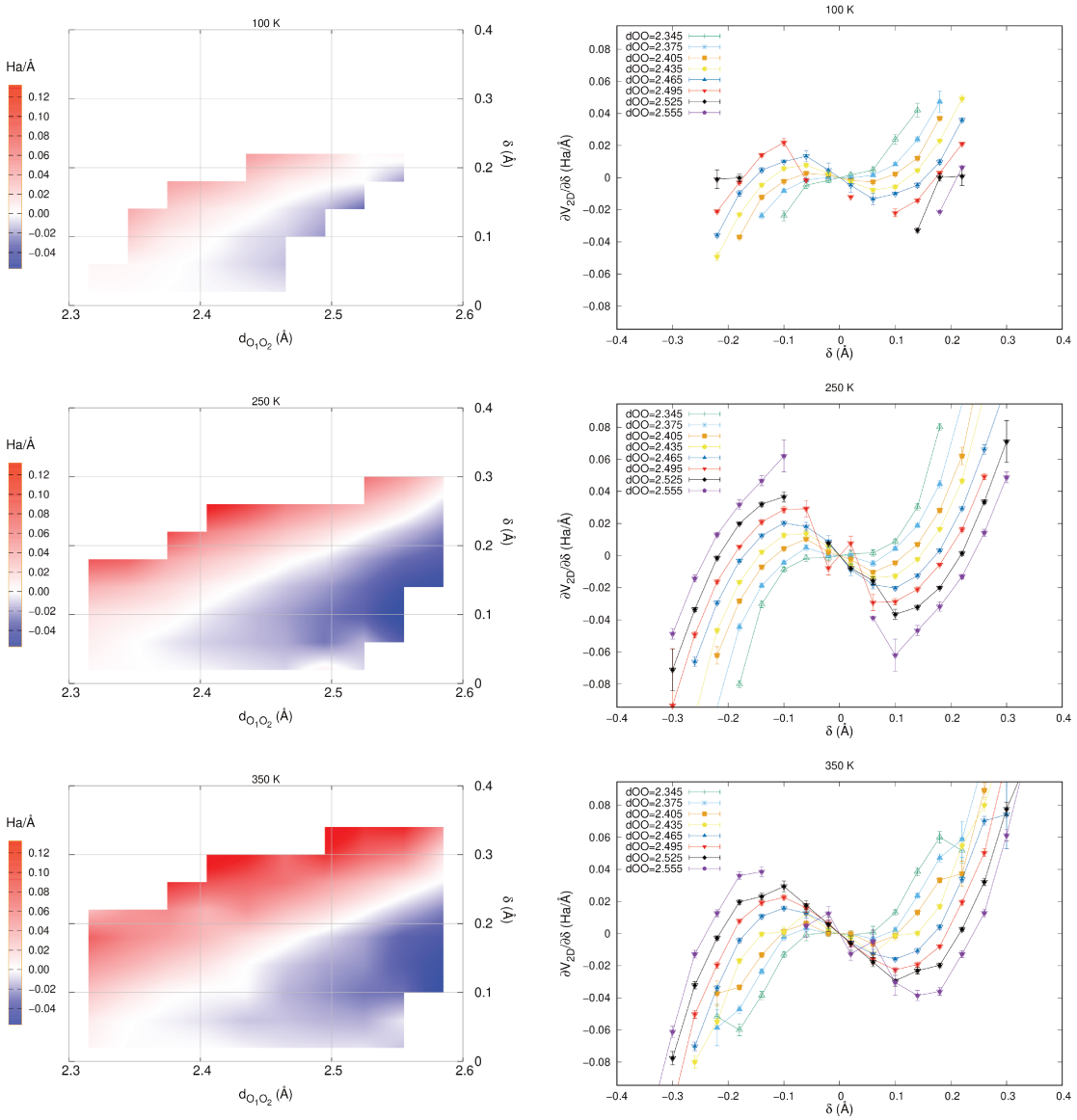


Figure B2: **Left column: contour plot of $\partial V_{2D}/\partial\delta_{H^+}$ as a function of both $d_{O_1O_2}$ and δ_{H^+}. Right column: superposition of $\partial V_{2D}/\partial\delta_{H^+}$, plotted as a function of δ_{H^+} at various (fixed) $d_{O_1O_2}$ values.** The force acting on H^+ projected along the O_1O_2 direction is given by $-\partial V_{2D}/\partial\delta_{H^+}$. Notice that the size of the $(d_{O_1O_2}, \delta_{H^+})$ space accessible by MD to sample these quantities increases as a function of the temperature.

ing the behavior of the $V_{2D}(d_{O_1O_2}, \delta_{H^+})$ potential not only around its equilibrium geometry but also over its tails. A way to overcome this issue within the projection framework described in Sec. B.1, is to sample the projected potential from QMC-MD simulations carried out at higher temperatures. As clearly shown in Fig. B2, at 250 K and 350 K the V_{2D} behavior can be evaluated on a much larger window in both δ_{H^+} and $d_{O_1O_2}$ directions. Moreover, we can increase the statistics of higher temperatures datapoints by averaging the 250 K and 350 K estimates.

We then fit the V_{1D} potential in Eq. B.5 by using a dataset averaged over 250 K and 350 K. The corresponding Morse potential fit is reported in Fig. B3b. From this analysis, the equilibrium

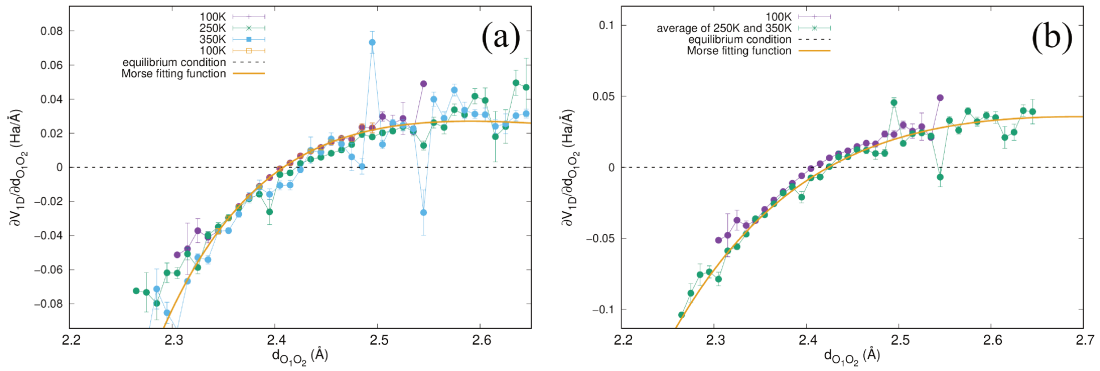


Figure B3: **Fit of the QMC estimates of the derivative of the Morse potential, $\partial V_{1D}/\partial d_{O_1O_2}$, as defined in Eq. B.5.** (a) Using the dataset of the trajectory at 100 K. (b) Using the dataset obtained by averaging the outcome of 250 K and 350 K simulations.

distance between two water molecules in the Zundel core of the protonated water hexamer is 2.425 Å at ≈ 300 K, again in a satisfactory agreement with the analysis based on the radial distribution function reported in Fig. 4.4 of the main text. Fitting over these datapoints leads to an increase of the equilibrium distance by 0.16 Å as the temperature is raised from 100 K to ≈ 300 K. The average based on the radial distribution function of the full VMC-MD simulations yields a cluster expansion of ≈ 0.25 Å.

The Morse potential determined from points computed at 100 K and the one from points averaged over 250 K and 350 K are plotted in Fig. B4, where the two fitting functions are superimposed. The ZPE analysis described in the main text of the thesis is carried out using the dataset averaged over 250 K and 350 K. As one can see in Fig. B4, in the energy range below 1000 K, the two curves are just shifted from one another, having nearly the same curvature around the minimum. Therefore, the conclusions on the ZPE effect reached by using a model potential projected at larger temperatures would not be different from the ones one could reach using the potential derived at 100 K.

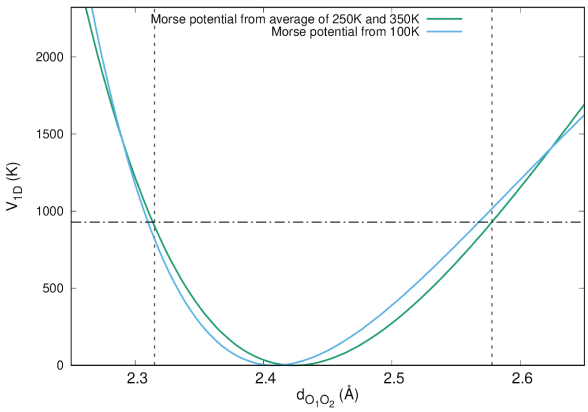


Figure B4: **V_{1D} determined from classical MD at** 100 K **and from the averaged dataset of classical MD at** 250 K **and** 350 K**.** The energy is expressed in Kelvin. The horizontal and vertical dashed lines are guides for the eye.

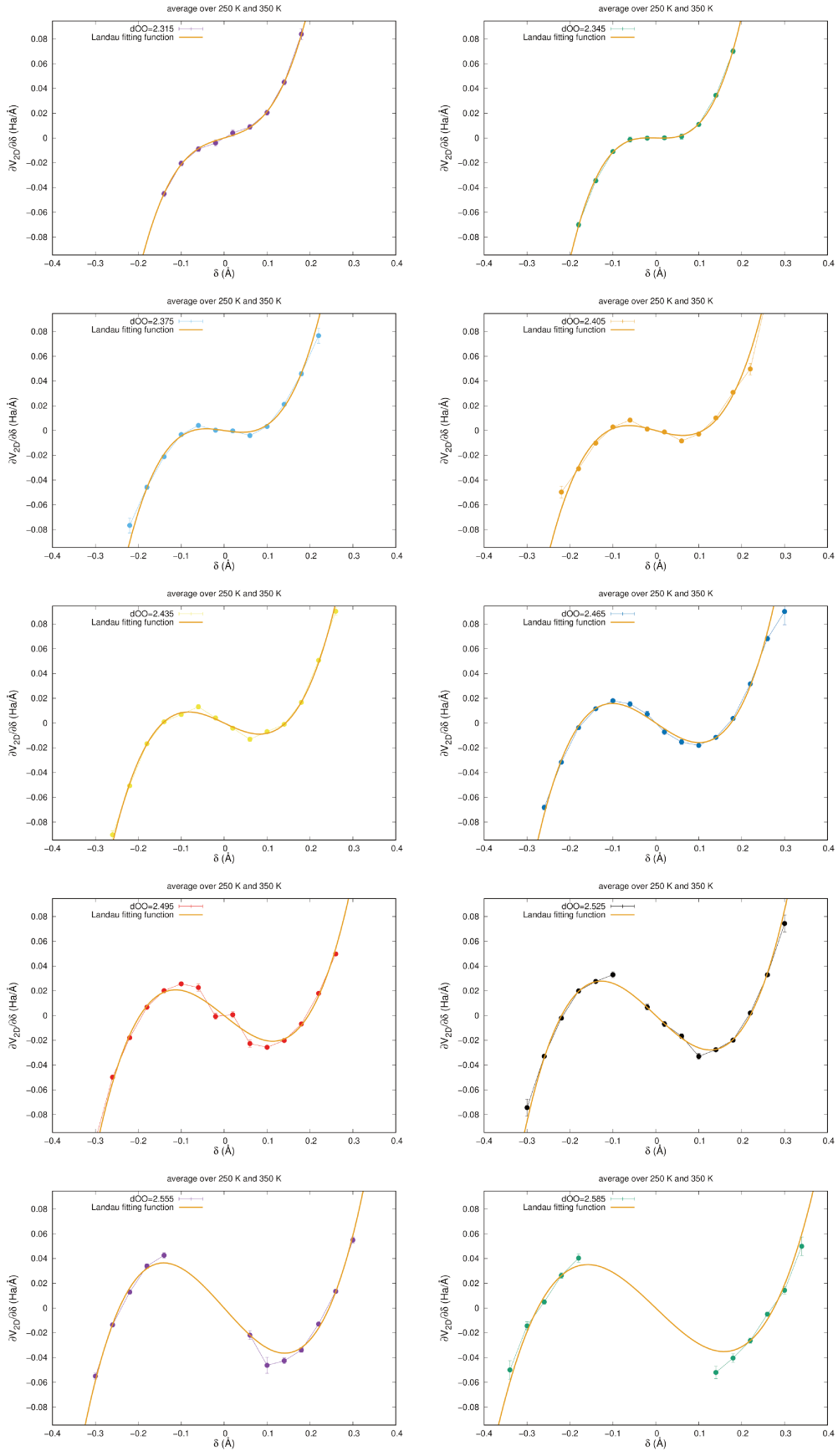


Figure B5: **Fit of the QMC forces using $\partial V_{d_{O_1O_2}}(x)/\partial x$ as fitting function for different $d_{O_1O_2}$, where the potential $V_{d_{O_1O_2}}(x)$ is defined in Eq. B.6.** The points are calculated as average over the two temperatures of 250 K and 350 K, reported in the right column of Fig. B2.

As a second step, we derive the $V_{2D}(d_{O_1O_2}, \delta_{H^+})$ potential. For every $d_{O_1O_2}$ slice, in Fig. B5 we plot the estimated values of the $\partial V_{2D}/\partial \delta_{H^+}$ derivative as a function of δ_{H^+}, computed by averaging over the 250 K and 350 K classical MD samples, as we did for the V_{1D} potential.

At every $d_{O_1O_2}$, we fit the derivative of the energy with respect to δ_{H^+}, by using a symmetric quartic function, i.e. a Landau potential, as fitting model for the energy dependence:

$$V_{d_{O_1O_2}}(x) = a + bx^2 + cx^4 \qquad \text{at fixed } d_{O_1O_2} \text{ distance.} \tag{B.6}$$

The fits for selected $d_{O_1O_2}$ values are also reported in Fig. B5. The parameters a, b and c have thus an implicit dependence on $d_{O_1O_2}$, which need to be further included in the full $V_{2D}(d_{O_1O_2}, \delta_{H^+})$ function. We found that a good parametrisation for b is given by:

$$b(d_{O_1O_2}) = \alpha + \beta d_{O_1O_2}, \tag{B.7}$$

while for c is:

$$c(d_{O_1O_2}) = \epsilon \exp(-\gamma d_{O_1O_2}). \tag{B.8}$$

Note that the $d_{O_1O_2}$-dependence in Eq. B.8 guarantees that the potential in Eq. B.6 always binds for $\epsilon > 0$. Fig. B6 demonstrates how this dependence, shown by the parameters evolution, is well taken into account by the functional forms in Eqs. B.7 and B.8.

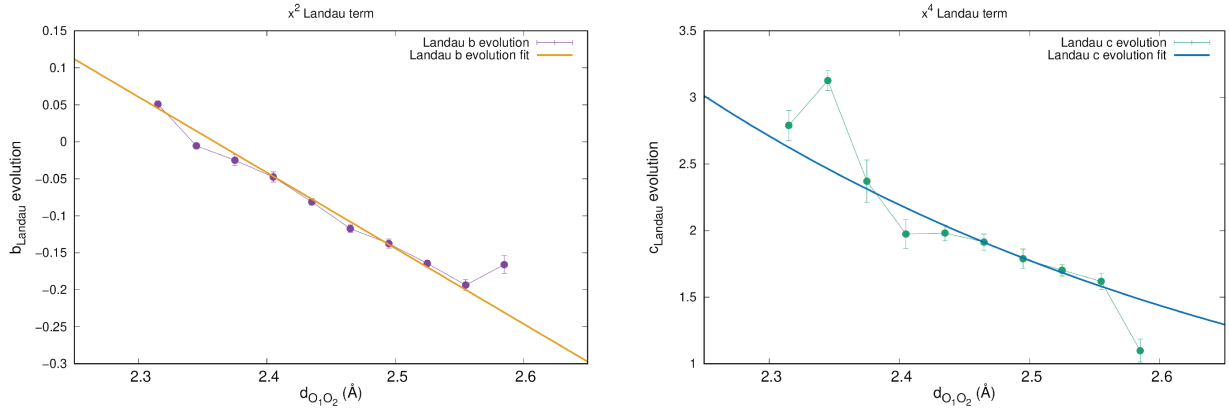


Figure B6: **Fit of the b and c dependence on the $d_{O_1O_2}$ distance, based on the functional forms in Eqs. B.7 and B.8.**

From Eq. B.7 and its fitting parameters, the bifurcation point turns out to be located at $d_{symm} \approx 2.37$ Å, in quite good agreement with the relaxation of the ground state geometry.

The final 2D potential $V_{2D}(d_{O_1O_2}, \delta_{H^+})$ is thus fully determined by the following function:

$$V_{2D}(d_{O_1O_2}, \delta_{H^+}) = a(d_{O_1O_2}) + b(d_{O_1O_2})\delta^2_{H^+} + c(d_{O_1O_2})\delta^4_{H^+}, \tag{B.9}$$

with $b(d_{O_1O_2})$ and $c(d_{O_1O_2})$ already defined in Eqs. B.7 and B.8, respectively, while $a(d_{O_1O_2})$ is defined as follows:

$$a(d_{O_1O_2}) = V_{1D}(d_{O_1O_2}) + \Delta(d_{O_1O_2}). \tag{B.10}$$

In the above Equation, $\Delta(d_{O_1O_2})$ is the proton barrier of the Landau potential $V_{d_{O_1O_2}}(x)$ in Eq. B.6, such that the bottom of $V_{2D}(d_{O_1O_2}, \delta_{H^+})$ at a given $d_{O_1O_2}$ distance follows exactly the Morse potential V_{1D} in Eq. B.5. In particular, $\Delta(d_{O_1O_2})$ reads:

$$\Delta(d_{O_1O_2}) = \begin{cases} 0, & \text{if } d_{O_1O_2} \leq d_{\text{symm}} \\ \frac{b^2(d_{O_1O_2})}{4c(d_{O_1O_2})}, & \text{otherwise.} \end{cases} \tag{B.11}$$

The resulting 2D potential $V_{2D}(d_{O_1O_2}, \delta_{H^+})$ and its derivative with respect to δ_{H^+} are drawn in the contour plot of Fig. B7. $\frac{\partial}{\partial \delta_{H^+}} V_{2D}(d_{O_1O_2}, \delta_{H^+})$ compares very well with the same quantity

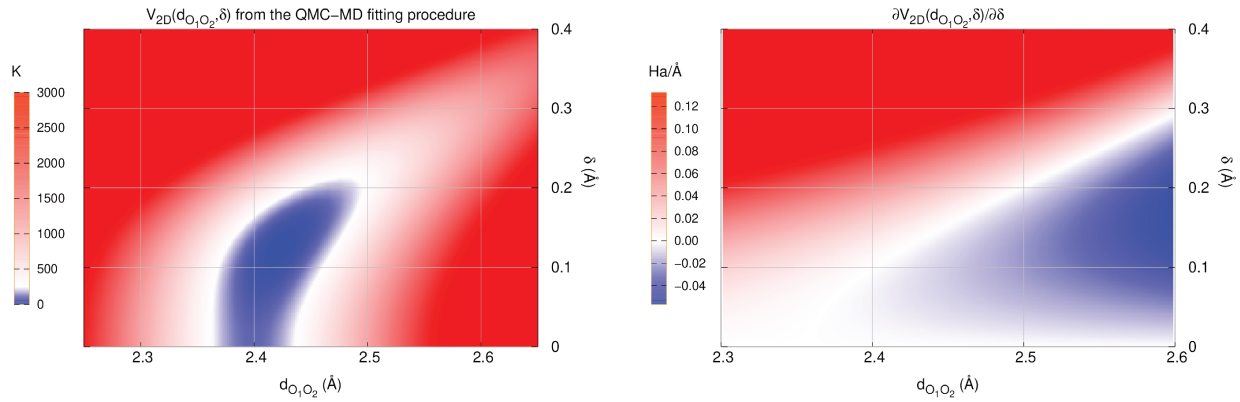


Figure B7: **Left panel: contour plot of the $V_{2D}(d_{O_1O_2}, \delta_{H^+})$ 2D model potential. Right panel: contour plot of the model-potential derivative $\frac{\partial}{\partial \delta_{H^+}} V_{2D}(d_{O_1O_2}, \delta_{H^+})$.** The latter can be compared with the contour plot of the mid- and bottom-left panels of Fig. B2, directly obtained from MD sampled datapoints.

directly evaluated by QMC-driven classical MD at both 250 K and 350 K, as shown in Fig. B2, mid- and bottom-left panels. This is an *a posteriori* check of the quality of our 2D-PES determination.

As we have seen, there is a residual temperature dependence in the determination of $V_{2D}(d_{O_1O_2}, \delta_{H^+})$, due to the projection scheme employed. Using a range of temperatures $T \in$ [250-350] K guarantees an optimal sampling of the configuration space during the MD, allowing for a more extended determination of the 2D-PES model. In Chapter 4, we used the function plotted in Fig. B7 to carry out an anharmonic vibrational analysis of the shuttling mode. The range of temperatures at which the model potential $V_{2D}(d_{O_1O_2}, \delta_{H^+})$ has been derived is consistent with the temperatures where the PT shows a "sweet spot", supporting the outcome of our analysis.

Appendix C

ML potentials hyper-parameters

C.1 OQML with FCHL19

OQML allows to fit simultaneously energies and forces leveraging on local atomic environments, which in our case are described using the FCHL19 representation.

As all descriptors, FCHL19 comes with some hyper-parameters, which are described in Section 6.2.2. These hyper-parameter can in principle be adapted to the specific dataset at hand, but in our case we used the default ones, already optimised on different and heterougenous datasets through Monte Carlo by the authors, as well as already tested on the Water40 dataset [278]. Their values are listed in the Table C.1

Hyper-parameter	E learning	$E + \mathbf{f}$ learning
n_{Rs2}	22	24
n_{Rs2}	17	22
r_{cut}	8.0	8.0
w	0.41	0.32
η_3	0.97	2.7
N_2	2.4	1.8
N_3	2.4	0.57
c_3	45.8	13.4
ς	π	π

Table C.1: **Hyper-parameters of the FCHL19 representation.** Depending on the type of labels that are learned, only energies or both energies and forces, they can be slightly different. From Ref. [279].

Also the regression method itself, OQML, comes with some hyper-parameters. Since there is not a closed-form expression of the optimisation problem, neither an iterative procedure, these hyper-parameters need to be tuned "by hand". This is usually done by algorithms that spans a portion of the grid of all the possible combination of hyper-parameters values, such as `GridSearch`.

C.2 MPNN with MACE

Hyper-parameter	Value
model	'MACE'
config_type_weights	'"Default":1.0'
E0s	'average'
r_max	6.0 Å
num_radial_basis	8
num_cutoff_basis	5
correlation	3
num_interactions	2
MLP_irreps	'16x0e'
radial_MLP	'[64, 64, 64]'
hidden_irreps	'128x0e + 128x1o'
num_channels	None
max_L	None
valid_fraction	0.1
loss	'weighted'
compute_stress	False
forces_weight	100.0
swa_forces_weight	10.0
energy_weight	1.0
swa_energy_weight	100.0
optimizer	'adam'
amsgrad	True
batch_size	20
valid_batch_size	1
lr	0.01
swa_lr	0.001
weight_decay	5e-07
scheduler	'ReduceLROnPlateau'
lr_factor	0.8
scheduler_patience	50
lr_scheduler_gamma	0.9993
swa	True
start_swa	80
ema	True
ema_decay	0.99
max_num_epochs	100
patience	2048
eval_interval	2

Table C.2: **MACE hyper-parameters.**

Bibliography

[1] C. J. D. T. Grotthuss. "Sur la décomposition de l'eau et des corps qu'elle tient en dissolution à l'aide de l'électricité Galvanique". In: *Annales de chimie* LVIII (1806), pp. 54–74.

[2] H. Danneel. "Notiz Über Ionengeschwindigkeiten". In: *Zeitschrift für Elektrochemie und angewandte physikalische Chemie* 11.16 (1905), pp. 249–252.

[3] I. M. Kolthoff. "The Confusion in the Expression of the So-Called "Hydrogen Ion Concentration" of a Solution and a Review of Brönsted's Conception of Acidity and Basicity". In: *Recueil des Travaux Chimiques des Pays-Bas* 49.5 (1930), pp. 401–414.

[4] J. D. Bernal and R. H. Fowler. "A Theory of Water and Ionic Solution, with Particular Reference to Hydrogen and Hydroxyl Ions". In: *The Journal of Chemical Physics* 1.8 (1933), pp. 515–548.

[5] G. Wannier. "Die Beweglichkeit Des Wasserstoff- Und Hydroxylions in Wäßriger Lösung. II". In: *Annalen der Physik* 416.7 (1935), pp. 569–590.

[6] R. P. Bell. *The Proton in Chemistry*. Springer US, 1973.

[7] C. E. Moore, B. Jaselskis, and A. von Smolinski. "The Proton". In: *Journal of Chemical Education* 62.10 (1985), p. 859.

[8] C. E. Moore, B. Jaselskis, and J. Florián. "Historical Development of the Hydrogen Ion Concept". In: *Journal of Chemical Education* 87.9 (2010), pp. 922–923.

[9] T. P. Silverstein. "The Solvated Proton Is NOT H_3O^+!" In: *Journal of Chemical Education* 88.7 (2011), pp. 875–875.

[10] S. Cukierman. "Et Tu, Grotthuss! And Other Unfinished Stories". In: *Biochimica et Biophysica Acta (BBA) - Bioenergetics*. Proton Transfer Reactions in Biological Systems 1757.8 (2006), pp. 876–885.

[11] D. Marx. "Proton Transfer 200 Years after von Grotthuss: Insights from Ab Initio Simulations". In: *ChemPhysChem* 7.9 (2006), pp. 1848–1870.

[12] N. Agmon, H. J. Bakker, R. K. Campen, R. H. Henchman, P. Pohl, S. Roke, M. Thämer, and A. Hassanali. "Protons and Hydroxide Ions in Aqueous Systems". In: *Chemical Reviews* 116.13 (2016), pp. 7642–7672.

[13] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Oxford University Press, 2017.

[14] N. Agmon. "The Grotthuss Mechanism". In: *Chemical Physics Letters* 244.5 (1995), pp. 456–462.

[15] C. Knight and G. A. Voth. "The Curious Case of the Hydrated Proton". In: *Accounts of Chemical Research* 45.1 (2012), pp. 101–109.

[16] A. Hassanali, F. Giberti, J. Cuny, T. D. Kühne, and M. Parrinello. "Proton Transfer through the Water Gossamer". In: *Proceedings of the National Academy of Sciences* 110.34 (2013), pp. 13723–13728.

[17] M. Eigen. "Proton Transfer, Acid-Base Catalysis, and Enzymatic Hydrolysis. Part I: Elementary Processes". In: *Angewandte Chemie International Edition in English* 3.1 (1964), pp. 1–19.

[18] G. Zundel and H. Metzger. "Energiebänder Der Tunnelnden Überschuß-Protonen in Flüssigen Säuren. Eine IR-spektroskopische Untersuchung Der Natur Der Gruppierungen H5O2+". In: *Zeitschrift für Physikalische Chemie* 58 (5_6 1968), pp. 225–245.

[19] J. A. Barker and R. O. Watts. "Structure of Water; A Monte Carlo Calculation". In: *Chemical Physics Letters* 3.3 (1969), pp. 144–145.

[20] A. Rahman and F. H. Stillinger. "Molecular Dynamics Study of Liquid Water". In: *The Journal of Chemical Physics* 55.7 (1971), pp. 3336–3359.

[21] M. D. Newton and S. Ehrenson. "Ab Initio Studies on the Structures and Energetics of Inner- and Outer-Shell Hydrates of the Proton and the Hydroxide Ion". In: *Journal of the American Chemical Society* 93.20 (1971), pp. 4971–4990.

[22] S. Scheiner. "Theoretical Studies of Proton Transfers". In: *Accounts of Chemical Research* 18.6 (1985), pp. 174–180.

[23] D. Borgis, G. Tarjus, and H. Azzouz. "An Adiabatic Dynamical Simulation Study of the Zundel Polarization of Strongly H-bonded Complexes in Solution". In: *The Journal of Chemical Physics* 97.2 (1992), pp. 1390–1400.

[24] J.-C. Jiang, Y.-S. Wang, H.-C. Chang, S. H. Lin, Y. T. Lee, G. Niedner-Schatteburg, and H.-C. Chang. "Infrared Spectra of H+(H2O)5-8 Clusters: Evidence for Symmetric Proton Hydration". In: *Journal of the American Chemical Society* 122.7 (2000), pp. 1398–1410.

[25] K. Mizuse and A. Fujii. "Infrared Photodissociation Spectroscopy of H+(H2O)6·Mm (M = Ne, Ar, Kr, Xe, H2, N2, and CH4): Messenger-Dependent Balance between H3O+ and H5O2+ Core Isomers". In: *Physical Chemistry Chemical Physics* 13.15 (2011), pp. 7129–7135.

[26] N. Heine, M. R. Fagiani, M. Rossi, T. Wende, G. Berden, V. Blum, and K. R. Asmis. "Isomer-Selective Detection of Hydrogen-Bond Vibrations in the Protonated Water Hexamer". In: *Journal of the American Chemical Society* 135.22 (2013), pp. 8266–8273.

[27] Z. Luz and S. Meiboom. "Rate and Mechanism of Proton Exchange in Aqueous Solutions of Phosphate Buffer". In: *Journal of the American Chemical Society* 86.22 (1964), pp. 4764–4766.

[28] D. M. Dennison and G. E. Uhlenbeck. "The Two-Minima Problem and the Ammonia Molecule". In: *Physical Review* 41.3 (1932), pp. 313–321.

[29] R. P. Bell. *The Tunnel Effect in Chemistry*. London ; New York: Chapman and Hall, 1980. 222 pp.

[30] F. Paesani and G. A. Voth. "The Properties of Water: Insights from Quantum Simulations". In: *The Journal of Physical Chemistry B* 113.17 (2009), pp. 5702–5719.

[31] D. Marx, A. Chandra, and M. E. Tuckerman. "Aqueous Basic Solutions: Hydroxide Solvation, Structural Diffusion, and Comparison to the Hydrated Proton". In: *Chem. Rev.* 110.4 (2010), pp. 2174–2216.

[32] M. Ceriotti, W. Fang, P. G. Kusalik, R. H. McKenzie, A. Michaelides, M. A. Morales, and T. E. Markland. "Nuclear Quantum Effects in Water and Aqueous Systems: Experiment, Theory, and Current Challenges". In: *Chemical Reviews* 116.13 (2016), pp. 7529–7550.

[33] E. Schrödinger. "Quantisierung Als Eigenwertproblem". In: *Annalen der Physik* 384.4 (1926), pp. 361–376.

[34] M. Born and R. Oppenheimer. "Zur Quantentheorie Der Molekeln". In: *Annalen der Physik* 389.20 (1927), pp. 457–484.

[35] T. Jecko. "On the Mathematical Treatment of the Born-Oppenheimer Approximation". In: *Journal of Mathematical Physics* 55.5 (2014), p. 053504.

[36] B. T. Sutcliffe and R. G. Woolley. "On the Quantum Theory of Molecules". In: *The Journal of Chemical Physics* 137.22 (2012), 22A544.

[37] M. Born. *Kopplung Der Elektronen- Und Kernbewegung in Molekeln Und Kristallen*. Göttingen: Vandenhoeck & Ruprecht, 1951.

[38] M. Born and K. Huang. *Dynamical Theory of Crystal Lattices*. International Series of Monographs on Physics. Clarendon press, 1954. 1 vol. (XII-420 p.) ; ill. ; 24 cm.

[39] D. Marx and J. Hutter. *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*. Cambridge University Press, 2009.

[40] J. C. Tully. "Mixed Quantum-Classical Dynamics: Mean-Field and Surface-Hopping". In: *Classical and Quantum Dynamics in Condensed Phase Simulations*. Lerici, Italy, 1998, pp. 489–514.

[41] R. Car and M. Parrinello. "Unified Approach for Molecular Dynamics and Density-Functional Theory". In: *Physical Review Letters* 55.22 (1985), pp. 2471–2474.

[42] J. C. Grossman, E. Schwegler, E. W. Draeger, F. Gygi, and G. Galli. "Towards an Assessment of the Accuracy of Density Functional Theory for First Principles Simulations of Water". In: *The Journal of Chemical Physics* 120.1 (2004), pp. 300–311.

[43] A. A. Hassanali, J. Cuny, V. Verdolino, and M. Parrinello. "Aqueous Solutions: State of the Art in Ab Initio Molecular Dynamics". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 372.2011 (2014), p. 20120482.

[44] W. Chen, F. Ambrosio, G. Miceli, and A. Pasquarello. "Ab Initio Electronic Structure of Liquid Water". In: *Physical Review Letters* 117.18 (2016), p. 186401.

[45] M. Chen, H.-Y. Ko, R. C. Remsing, M. F. Calegari Andrade, B. Santra, Z. Sun, A. Selloni, R. Car, M. L. Klein, J. P. Perdew, and X. Wu. "Ab Initio Theory and Modeling of Water". In: *Proceedings of the National Academy of Sciences* 114.41 (2017), pp. 10846–10851.

[46] M. E. Tuckerman, K. Laasonen, M. Sprik, and M. Parrinello. "Ab Initio Simulations of Water and Water Ions". In: *Journal of Physics: Condensed Matter* 6 (23A 1994), A93.

[47] M. Tuckerman, K. Laasonen, M. Sprik, and M. Parrinello. "Ab Initio Molecular Dynamics Simulation of the Solvation and Transport of Hydronium and Hydroxyl Ions in Water". In: *The Journal of Chemical Physics* 103.1 (1995), pp. 150–161.

[48] D. Marx, M. E. Tuckerman, J. Hutter, and M. Parrinello. "The Nature of the Hydrated Excess Proton in Water". In: *Nature* 397.6720 (1999), pp. 601–604.

[49] D. Marx, M. E. Tuckerman, and M. Parrinello. "Solvated Excess Protons in Water: Quantum Effects on the Hydration Structure". In: *Journal of Physics: Condensed Matter* 12 (8A 2000), A153–A159.

[50] D. Asthagiri, L. R. Pratt, and J. D. Kress. "Ab Initio Molecular Dynamics and Quasichemical Study of H+(Aq)". In: *Proceedings of the National Academy of Sciences* 102.19 (2005), pp. 6704–6708.

[51] A. Bodi, J. Csontos, M. Kállay, S. Borkar, and B. Sztáray. "On the Protonation of Water". In: *Chemical Science* 5.8 (2014), pp. 3057–3063.

[52] O. Vendrell, F. Gatti, and H.-D. Meyer. "Full Dimensional (15-Dimensional) Quantum-Dynamical Simulation of the Protonated Water Dimer. II. Infrared Spectrum and Vibrational Dynamics". In: *The Journal of Chemical Physics* 127.18 (2007), p. 184303.

[53] K. Yuan, Y. Cheng, L. Cheng, Q. Guo, D. Dai, X. Wang, X. Yang, and R. N. Dixon. "Nonadiabatic Dissociation Dynamics in H2O: Competition between Rotationally and Nonrotationally Mediated Pathways". In: *Proceedings of the National Academy of Sciences* 105.49 (2008), pp. 19148–19153.

[54] Z. P. Wang, P. M. Dinh, P. G. Reinhard, and E. Suraud. "Ultrafast Nonadiabatic Dynamics of a Water Dimer in Femtosecond Laser Pulses". In: *Laser Physics* 24.10 (2014), p. 106004.

[55] L. Lu, A. Wildman, A. J. Jenkins, L. Young, A. E. Clark, and X. Li. "The "Hole" Story in Ionized Water from the Perspective of Ehrenfest Dynamics". In: *The Journal of Physical Chemistry Letters* 11.22 (2020), pp. 9946–9951.

[56] K. Drukker, S. W. de Leeuw, and S. Hammes-Schiffer. "Proton Transport along Water Chains in an Electric Field". In: *The Journal of Chemical Physics* 108.16 (1998), pp. 6799–6808.

[57] V. Sharma and M. Fernández-Serra. "Proton-Transfer Dynamics in Ionized Water Chains Using Real-Time Time-Dependent Density Functional Theory". In: *Physical Review Research* 2.4 (2020), p. 043082.

[58] K. Schnorr, M. Belina, S. Augustin, H. Lindenblatt, Y. Liu, S. Meister, T. Pfeifer, G. Schmid, R. Treusch, F. Trost, P. Slavíček, and R. Moshammer. "Direct Tracking of Ultrafast Proton Transfer in Water Dimers". In: *Science Advances* 9.28 (2023), eadg7864.

[59] J. W. S. Rayleigh. *The Theory of Sound*. London : Macmillan, 1894–1896. 534 pp.

[60] W. Ritz. "Über eine neue Methode zur Lösung gewisser Variationsprobleme der mathematischen Physik." In: *Journal für die reine und angewandte Mathematik* 135 (1909), pp. 1–61.

[61] A. Szabo and N. S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Mineola, N.Y: Dover Publications, 1996. 466 pp.

[62] J. C. Slater. "Atomic Shielding Constants". In: *Physical Review* 36.1 (1930), pp. 57–64.

[63] S. F. Boys and A. C. Egerton. "Electronic Wave Functions - I. A General Method of Calculation for the Stationary States of Any Molecular System". In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 200.1063 (1997), pp. 542–554.

[64] D. R. Hartree. "The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 24.1 (1928), pp. 89–110.

[65] V. Fock. "Naherungsmethode zur Losung des quantenmechanischen Mehrkorperproblems". In: *Zeitschrift f□r Physik* 61.1-2 (1930), pp. 126–148.

[66] J. C. Slater. "The Theory of Complex Spectra". In: *Physical Review* 34.10 (1929), pp. 1293–1322.

[67] Chr. Møller and M. S. Plesset. "Note on an Approximation Treatment for Many-Electron Systems". In: *Physical Review* 46.7 (1934), pp. 618–622.

[68] F. Coester. "Bound States of a Many-Particle System". In: *Nuclear Physics* 7 (1958), pp. 421–424.

[69] F. Coester and H. Kümmel. "Short-Range Correlations in Nuclear Wave Functions". In: *Nuclear Physics* 17 (1960), pp. 477–485.

[70] J. Čížek. "On the Correlation Problem in Atomic and Molecular Systems. Calculation of Wavefunction Components in Ursell-Type Expansion Using Quantum-Field Theoretical Methods". In: *The Journal of Chemical Physics* 45.11 (1966), pp. 4256–4266.

[71] J. Paldus and X. Li. "A Critical Assessment of Coupled Cluster Method in Quantum Chemistry". In: *Advances in Chemical Physics*. John Wiley & Sons, Ltd, 1999, pp. 1–175.

[72] P. Hohenberg and W. Kohn. "Inhomogeneous Electron Gas". In: *Phys. Rev. B* 136 (3B 1964), pp. 864–871.

[73] R. M. Martin. *Electronic Structure: Basic Theory and Practical Methods*. Second edition. Cambridge, United Kingdom ; New York, NY: Cambridge University Press, 2020.

[74] W. Kohn and L. J. Sham. "Self-Consistent Equations Including Exchange and Correlation Effects". In: *Phys. Rev.* 140 (4A 1965), A1133–A1138.

[75] D. M. Ceperley and B. J. Alder. "Ground State of the Electron Gas by a Stochastic Method". In: *Phys. Rev. Lett.* 45.7 (1980), pp. 566–569.

[76] J. P. Perdew and A. Zunger. "Self-Interaction Correction to Density-Functional Approximations for Many-Electron Systems". In: *Phys. Rev. B* 23.10 (1981), pp. 5048–5079.

[77] J. P. Perdew, K. Burke, and M. Ernzerhof. "Generalized Gradient Approximation Made Simple". In: *Physical Review Letters* 77.18 (1996), pp. 3865–3868.

[78] A. D. Becke. "Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior". In: *Phys. Rev. A* 38.6 (1988), pp. 3098–3100.

[79] C. Lee, W. Yang, and R. G. Parr. "Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density". In: *Phys. Rev. B* 37.2 (1988), pp. 785–789.

[80] C. Lee, H. Chen, and G. Fitzgerald. "Structures of the Water Hexamer Using Density Functional Methods". In: *The Journal of Chemical Physics* 101.5 (1994), pp. 4472–4473.

[81] C. Lee, H. Chen, and G. Fitzgerald. "Chemical Bonding in Water Clusters". In: *The Journal of Chemical Physics* 102.3 (1995), pp. 1266–1269.

[82] J. Harris. "Simplified Method for Calculating the Energy of Weakly Interacting Fragments". In: *Phys. Rev. B* 31.4 (1985), pp. 1770–1779.

[83] M. J. McGrath, J. I. Siepmann, I.-F. W. Kuo, C. J. Mundy, J. VandeVondele, J. Hutter, F. Mohamed, and M. Krack. "Isobaric–Isothermal Monte Carlo Simulations from First Principles: Application to Liquid Water at Ambient Conditions". In: *ChemPhysChem* 6.9 (2005), pp. 1894–1901.

[84] J. Schmidt, J. VandeVondele, I.-F. W. Kuo, D. Sebastiani, J. I. Siepmann, J. Hutter, and C. J. Mundy. "Isobaric−Isothermal Molecular Dynamics Simulations Utilizing Density Functional Theory: An Assessment of the Structure and Density of Water at Near-Ambient Conditions". In: *The Journal of Physical Chemistry B* 113.35 (2009), pp. 11959–11964.

[85] B. Santra, A. Michaelides, and M. Scheffler. "Coupled Cluster Benchmarks of Water Monomers and Dimers Extracted from Density-Functional Theory Liquid Water: The Importance of Monomer Deformations". In: *The Journal of Chemical Physics* 131.12 (2009), p. 124509.

[86] S. Grimme. "Semiempirical GGA-type Density Functional Constructed with a Long-Range Dispersion Correction". In: *J. Comput. Chem.* 27.15 (2006), pp. 1787–1799.

[87] A. D. Becke. "On the Large-gradient Behavior of the Density Functional Exchange Energy". In: *The Journal of Chemical Physics* 85.12 (1986), pp. 7184–7187.

[88] J. P. Perdew and Y. Wang. "Accurate and Simple Analytic Representation of the Electron-Gas Correlation Energy". In: *Physical Review B* 45.23 (1992), pp. 13244–13249.

[89] K. Lee, É. D. Murray, L. Kong, B. I. Lundqvist, and D. C. Langreth. "Higher-Accuracy van Der Waals Density Functional". In: *Phys. Rev. B* 82.8 (2010).

[90] M. Dion, H. Rydberg, E. Schröder, D. C. Langreth, and B. I. Lundqvist. "Van Der Waals Density Functional for General Geometries". In: *Phys. Rev. Lett.* 92.24 (2004), p. 175.

[91] A. K. Kelkkanen, B. I. Lundqvist, and J. K. Nørskov. "Density Functional for van Der Waals Forces Accounts for Hydrogen Bond in Benchmark Set of Water Hexamers". In: *The Journal of Chemical Physics* 131.4 (2009), p. 046102.

[92] A. Møgelhøj, A. K. Kelkkanen, K. T. Wikfeldt, J. Schiøtz, J. J. Mortensen, L. G. M. Pettersson, B. I. Lundqvist, K. W. Jacobsen, A. Nilsson, and J. K. Nørskov. "Ab Initio van Der Waals Interactions in Simulations of Water Alter Structure from Mainly Tetrahedral to High-Density-Like". In: *J. Phys. Chem. B* 115.48 (2011), pp. 14149–14160.

[93] C. Zhang, J. Wu, G. Galli, and F. Gygi. "Structural and Vibrational Properties of Liquid Water from van Der Waals Density Functionals". In: *J. Chem. Theory Comput.* 7.10 (2011), pp. 3054–3061.

[94] J. Toulouse, R. Assaraf, and C. J. Umrigar. "Chapter Fifteen - Introduction to the Variational and Diffusion Monte Carlo Methods". In: *Electron Correlation in Molecules Ab Initio Beyond Gaussian Quantum Chemistry*. Ed. by P. E. Hoggan and T. Ozdogan. Vol. 73. Advances in Quantum Chemistry. Academic Press, 2016, pp. 285–314.

[95] F. Becca and S. Sorella. *Quantum Monte Carlo Approaches for Correlated Systems*. Cambridge: Cambridge University Press, 2017.

[96] W. M. C. Foulkes, L. Mitas, R. J. Needs, and G. Rajagopal. "Quantum Monte Carlo Simulations of Solids". In: *Rev. Mod. Phys.* 73.1 (2001), pp. 33–83.

[97] B. M. Austin, D. Y. Zubarev, and W. A. Lester. "Quantum Monte Carlo and Related Approaches". In: *Chemical Reviews* 112.1 (2012), pp. 263–288.

[98] W. L. McMillan. "Ground State of Liquid ${\mathrm{He}}^{4}$". In: *Physical Review* 138 (2A 1965), A442–A451.

[99] D. Ceperley, G. V. Chester, and M. H. Kalos. "Monte Carlo Simulation of a Many-Fermion Study". In: *Physical Review B* 16.7 (1977), pp. 3081–3099.

[100] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. "Equation of State Calculations by Fast Computing Machines". In: *The Journal of Chemical Physics* 21.6 (1953), pp. 1087–1092.

[101] W. K. Hastings. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications". In: *Biometrika* 57.1 (1970), pp. 97–109.

[102] C. J. Umrigar. "Accelerated Metropolis Method". In: *Physical Review Letters* 71.3 (1993), pp. 408–411.

[103] H. Flyvbjerg and H. G. Petersen. "Error Estimates on Averages of Correlated Data". In: *The Journal of Chemical Physics* 91.1 (1989), pp. 461–466.

[104] R. E. Lowther and R. L. Coldwell. "Monte Carlo Calculation of the Born-Oppenheimer Potential between Two Helium Atoms". In: *Physical Review A* 22.1 (1980), pp. 14–21.

[105] C. J. Umrigar. "Two Aspects of Quantum Monte Carlo: Determination of Accurate Wavefunctions and Determination of Potential Energy Surfaces of Molecules". In: *International Journal of Quantum Chemistry* 36.S23 (1989), pp. 217–230.

[106] C. Filippi and C. J. Umrigar. "Correlated Sampling in Quantum Monte Carlo: A Route to Forces". In: *Phys. Rev. B* 61.24 (2000), R16291–R16294.

[107] S. Sorella and L. Capriotti. "Algorithmic Differentiation and the Calculation of Forces by Quantum Monte Carlo". In: *J. Chem. Phys.* 133.23 (2010), p. 234111.

[108] K. Nakano, A. Raghav, and S. Sorella. "Space-Warp Coordinate Transformation for Efficient Ionic Force Calculations in Quantum Monte Carlo". In: *The Journal of Chemical Physics* 156.3 (2022), p. 034101.

[109] K. Nakano, M. Casula, and G. Tenti. "Efficient Calculation of Unbiased Atomic Forces in Ab Initio Variational Monte Carlo". In: *Physical Review B* 109.20 (2024), p. 205151.

[110] R. Assaraf and M. Caffarel. "Computing Forces with Quantum Monte Carlo". In: *The Journal of Chemical Physics* 113.10 (2000), pp. 4028–4034.

[111] R. Assaraf and M. Caffarel. "Zero-Variance Zero-Bias Principle for Observables in Quantum Monte Carlo: Application to Forces". In: *J. Chem. Phys.* 119.20 (2003), pp. 10536–10552.

[112] C. Attaccalite and S. Sorella. "Stable Liquid Hydrogen at High Pressure by a Novel Ab Initio Molecular-Dynamics Calculation". In: *Physical Review Letters* 100.11 (2008), p. 114501.

[113] A. Zen, Y. Luo, S. Sorella, and L. Guidoni. "Molecular Properties by Quantum Monte Carlo: An Investigation on the Role of the Wave Function Ansatz and the Basis Set in the Water Molecule". In: *Journal of Chemical Theory and Computation* 9.10 (2013), pp. 4332–4350.

[114] M. Ogata and A. Himeda. *Effect of Exclusion of Double Occupancies in T-J Model: Extension of Gutzwiller Approximation*. 2000. URL: http://arxiv.org/abs/cond-mat/0003465. Pre-published.

[115] S.-i. Amari. "Natural Gradient Works Efficiently in Learning". In: *Neural Computation* 10.2 (1998), pp. 251–276.

[116] S. Sorella. "Green Function Monte Carlo with Stochastic Reconfiguration". In: *Physical Review Letters* 80.20 (1998), pp. 4558–4561.

[117] M. Casula, C. Attaccalite, and S. Sorella. "Correlated Geminal Wave Function for Molecules: An Efficient Resonating Valence Bond Approach". In: *The Journal of Chemical Physics* 121.15 (2004), pp. 7110–7126.

[118] C. J. Umrigar, J. Toulouse, C. Filippi, S. Sorella, and R. G. Hennig. "Alleviation of the Fermion-Sign Problem by Optimization of Many-Body Wave Functions". In: *Physical Review Letters* 98.11 (2007), p. 110201.

[119] M. Casula and S. Sorella. "Geminal Wave Functions with Jastrow Correlation: A First Application to Atoms". In: *The Journal of Chemical Physics* 119.13 (2003), pp. 6500–6511.

[120] N. Dupuy, S. Bouaouli, F. Mauri, S. Sorella, and M. Casula. "Vertical and Adiabatic Excitations in Anthracene from Quantum Monte Carlo: Constrained Energy Minimization for Structural and Electronic Excited-State Properties in the JAGP Ansatz". In: *The Journal of Chemical Physics* 142.21 (2015), p. 214109.

[121] R. Jastrow. "Many-Body Problem with Strong Forces". In: *Physical Review* 98.5 (1955), pp. 1479–1484.

[122] L. Pauling. *The Nature of the Chemical Bond and the Structure of Molecules and Crystals: An Introduction to Modern Structural Chemistry*. Ithaca, New York: Cornell University Press, 1939.

[123] P. W. Anderson. "Resonating Valence Bonds: A New Kind of Insulator?" In: *Materials Research Bulletin* 8.2 (1973), pp. 153–160.

[124] S. Sorella, M. Casula, and D. Rocca. "Weak Binding between Two Aromatic Rings: Feeling the van Der Waals Attraction by Quantum Monte Carlo Methods". In: *J. Chem. Phys.* 127.1 (2007), p. 014105.

[125] T. Kato. "On the Eigenfunctions of Many-Particle Systems in Quantum Mechanics". In: *Communications on Pure and Applied Mathematics* 10.2 (1957), pp. 151–177.

[126] M. Burkatzki, C. Filippi, and M. Dolg. "Energy-Consistent Pseudopotentials for Quantum Monte Carlo Calculations". In: *J. Chem. Phys.* 126.23 (2007), p. 234105.

[127] F. Mouhat, M. Peria, T. Morresi, R. Vuilleumier, A. M. Saitta, and M. Casula. "Thermal Dependence of the Hydrated Proton and Optimal Proton Transfer in the Protonated Water Hexamer". In: *Nature Communications* 14.1 (2023), p. 6930.

[128] S. Sorella, N. Devaux, M. Dagrada, G. Mazzola, and M. Casula. "Geminal Embedding Scheme for Optimal Atomic Basis Set Construction in Correlated Calculations". In: *J. Chem. Phys.* 143.24 (2015), p. 244112.

[129] S. Sorella. "Generalized Lanczos Algorithm for Variational Quantum Monte Carlo". In: *Phys. Rev. B* 64.2 (2001), p. 024512.

[130] M. Dagrada, M. Casula, A. M. Saitta, S. Sorella, and F. Mauri. "Quantum Monte Carlo Study of the Protonated Water Dimer". In: *Journal of Chemical Theory and Computation* 10.5 (2014), pp. 1980–1993.

[131] F. Mouhat, S. Sorella, R. Vuilleumier, A. M. Saitta, and M. Casula. "Fully Quantum Description of the Zundel Ion: Combining Variational Quantum Monte Carlo with Path Integral Langevin Dynamics". In: *Journal of Chemical Theory and Computation* 13.6 (2017), pp. 2400–2417.

[132] E. Neuscamman. "Size Consistency Error in the Antisymmetric Geminal Power Wave Function Can Be Completely Removed". In: *Physical Review Letters* 109.20 (2012), p. 203001.

[133] F. Mouhat. "Fully Quantum Dynamics of Protonated Water Clusters". These de doctorat. Sorbonne université, 2018.

[134] R. P. Feynman. "Forces in Molecules". In: *Physical Review* 56.4 (1939), pp. 340–343.

[135] L. Verlet. "Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules". In: *Physical Review* 159.1 (1967), pp. 98–103.

[136] W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson. "A Computer Simulation Method for the Calculation of Equilibrium Constants for the Formation of Physical Clusters of Molecules: Application to Small Water Clusters". In: *The Journal of Chemical Physics* 76.1 (1982), pp. 637–649.

[137] H. F. Trotter. "On the Product of Semi-Groups of Operators". In: *Proceedings of the American Mathematical Society* 10.4 (1959), pp. 545–551.

[138] M. Suzuki. "Decomposition Formulas of Exponential Operators and Lie Exponentials with Some Applications to Quantum Mechanics and Statistical Physics". In: *Journal of Mathematical Physics* 26.4 (1985), pp. 601–612.

[139] S. Nosé. "A Unified Formulation of the Constant Temperature Molecular Dynamics Methods". In: *The Journal of Chemical Physics* 81.1 (1984), pp. 511–519.

[140] W. G. Hoover. "Canonical Dynamics: Equilibrium Phase-Space Distributions". In: *Physical Review A* 31.3 (1985), pp. 1695–1697.

[141] L. C. Evans. *An Introduction to Stochastic Differential Equations*. Providence, RI: American Mathematical Society, 2013. 151 pp.

[142] H. Risken and T. Frank. *The Fokker-Planck Equation: Methods of Solution and Applications*. 2nd ed. Springer Series in Synergetics. Berlin Heidelberg: Springer-Verlag, 1996.

[143] R. Kubo. "The Fluctuation-Dissipation Theorem". In: *Reports on Progress in Physics* 29.1 (1966), pp. 255–284.

[144] G. E. Uhlenbeck and L. S. Ornstein. "On the Theory of the Brownian Motion". In: *Physical Review* 36.5 (1930), pp. 823–841.

[145] G. Bussi and M. Parrinello. "Accurate Sampling Using Langevin Dynamics". In: *Physical Review E* 75.5 (2007), p. 056707.

[146] D. S. Lemons and P. Langevin. *An Introduction to Stochastic Processes in Physics: Containing "On the Theory of Brownian Motion" by Paul Langevin, Translated by Anthony Gythiel*. Baltimore: Johns Hopkins University Press, 2002. 110 pp.

[147] Y. Luo, A. Zen, and S. Sorella. "Ab Initio Molecular Dynamics with Noisy Forces: Validating the Quantum Monte Carlo Approach with Benchmark Calculations of Molecular Vibrational Properties". In: *The Journal of Chemical Physics* 141.19 (2014), p. 194112.

[148] G. Mazzola, S. Yunoki, and S. Sorella. "Unexpectedly High Pressure for Molecular Dissociation in Liquid Hydrogen by Electronic Simulation". In: *Nature Communications* 5.1 (2014), p. 3487.

[149] A. Zen, Y. Luo, G. Mazzola, L. Guidoni, and S. Sorella. "Ab Initio Molecular Dynamics Simulation of Liquid Water by Quantum Monte Carlo". In: *The Journal of Chemical Physics* 142.14 (2015), p. 144111.

[150] F. Hund. "Zur Deutung der Molekelspektren. III." In: *Zeitschrift für Physik* 43.11 (1927), pp. 805–826.

[151] J. Waluk. "Nuclear Quantum Effects in Proton or Hydrogen Transfer". In: *The Journal of Physical Chemistry Letters* (2024).

[152] M. A. Lill and V. Helms. "Reaction Rates for Proton Transfer over Small Barriers and Connection to Transition State Theory". In: *The Journal of Chemical Physics* 115.17 (2001), pp. 7985–7992.

[153] R. P. Feynman and A. R. Hibbs. *Quantum Mechanics and Path Integrals*. New York: McGraw-Hill, 1965.

[154] D. Chandler and P. G. Wolynes. "Exploiting the Isomorphism between Quantum Theory and Classical Statistical Mechanics of Polyatomic Fluids". In: *The Journal of Chemical Physics* 74.7 (1981), pp. 4078–4095.

[155] M. Sprik, M. L. Klein, and D. Chandler. "Staging: A Sampling Technique for the Monte Carlo Evaluation of Path Integrals". In: *Physical Review B* 31.7 (1985), pp. 4234–4244.

[156] T. E. Markland and M. Ceriotti. "Nuclear Quantum Effects Enter the Mainstream". In: *Nature Reviews Chemistry* 2.3 (3 2018), pp. 1–14.

[157] R. A. Kuharski and P. J. Rossky. "Quantum Mechanical Contributions to the Structure of Liquid Water". In: *Chemical Physics Letters* 103.5 (1984), pp. 357–362.

[158] A. Wallqvist and B. J. Berne. "Path-Integral Simulation of Pure Water". In: *Chemical Physics Letters* 117.3 (1985), pp. 214–219.

[159] M. Ceriotti, J. Cuny, M. Parrinello, and D. E. Manolopoulos. "Nuclear Quantum Effects and Hydrogen Bond Fluctuations in Water". In: *Proceedings of the National Academy of Sciences* 110.39 (2013), pp. 15591–15596.

[160] K. Suzuki, M. Shiga, and M. Tachikawa. "Temperature and Isotope Effects on Water Cluster Ions with Path Integral Molecular Dynamics Based on the Fourth Order Trotter Expansion". In: *The Journal of Chemical Physics* 129.14 (2008), p. 144310.

[161] K. Suzuki, M. Tachikawa, and M. Shiga. "Temperature Dependence on the Structure of Zundel Cation and Its Isotopomers". In: *J. Chem. Phys.* 138.18 (2013), p. 184307.

[162] M. E. Tuckerman, D. Marx, M. L. Klein, and M. Parrinello. "On the Quantum Nature of the Shared Proton in Hydrogen Bonds". In: *Science* 275.5301 (1997), pp. 817–820.

[163] M. Benoit, D. Marx, and M. Parrinello. "Tunnelling and Zero-Point Motion in High-Pressure Ice". In: *Nature* (1998).

[164] N. Pugliano. "Vibration-Rotation-Tunneling Dynamics in Small Water Clusters". Lawrence Berkeley Lab., CA (United States), 1992.

[165] N. Pugliano and R. J. Saykally. "Measurement of Quantum Tunneling Between Chiral Isomers of the Cyclic Water Trimer". In: *Science* 257.5078 (1992), pp. 1937–1940.

[166] K. R. Liedl, S. Sekušak, R. T. Kroemer, and B. M. Rode. "New Insights into the Dynamics of Concerted Proton Tunneling in Cyclic Water and Hydrogen Fluoride Clusters". In: *The Journal of Physical Chemistry A* 101.26 (1997), pp. 4707–4716.

[167] M. E. Tuckerman and A. Hughes. "Path Integral Molecular Dynamics: A Computational Approach to Quantum Statistical Mechanics". In: *Classical and Quantum Dynamics in Condensed Phase Simulations*. WORLD SCIENTIFIC, 1998, pp. 311–357.

[168] R. W. Hall and B. J. Berne. "Nonergodicity in Path Integral Molecular Dynamics". In: *The Journal of Chemical Physics* 81.8 (1984), pp. 3641–3643.

[169] G. J. Martyna, M. L. Klein, and M. Tuckerman. "Nosé–Hoover Chains: The Canonical Ensemble via Continuous Dynamics". In: *The Journal of Chemical Physics* 97.4 (1992), pp. 2635–2643.

[170] M. Ceriotti, M. Parrinello, T. E. Markland, and D. E. Manolopoulos. "Efficient Stochastic Thermostatting of Path Integral Molecular Dynamics". In: *The Journal of Chemical Physics* 133.12 (2010), p. 124104.

[171] T. E. Markland and D. E. Manolopoulos. "A Refined Ring Polymer Contraction Scheme for Systems with Electrostatic Interactions". In: *Chemical Physics Letters* 464.4-6 (2008), pp. 256–261.

[172] T. E. Markland and D. E. Manolopoulos. "An Efficient Ring Polymer Contraction Scheme for Imaginary Time Path Integral Simulations". In: *J. Chem. Phys.* 129.2 (2008), p. 024105.

[173] M. Ceriotti, D. E. Manolopoulos, and M. Parrinello. "Accelerating the Convergence of Path Integral Dynamics with a Generalized Langevin Equation". In: *J. Chem. Phys.* 134.8 (2011), p. 084104.

[174] M. Mella, J.-L. Kuo, D. C. Clary, and M. L. Klein. "Nuclear Quantum Effects on the Structure and Energetics of (H2O)6H+". In: *Physical Chemistry Chemical Physics* 7.11 (2005), p. 2324.

[175] J. P. Heindel, Q. Yu, J. M. Bowman, and S. S. Xantheas. "Benchmark Electronic Structure Calculations for H□O□ (H□O)_n, n= 0-5, Clusters and Tests of an Existing 1, 2, 3-Body Potential Energy Surface with a New 4-Body Correction". In: *Journal of Chemical Theory and Computation* 14.9 (2018), pp. 4553–4566.

[176] J. M. Finney, T. H. Choi, R. M. Huchmala, J. P. Heindel, S. S. Xantheas, K. D. Jordan, and A. B. McCoy. "Isotope Effects in the Zundel–Eigen Isomerization of H+(H2O)6". In: *The Journal of Physical Chemistry Letters* 14.20 (2023), pp. 4666–4672.

[177] D. J. Jones, J. Rozière, J. Penfold, and J. Tomkinson. "Incoherent Inelastic Neutron Scattering Studies of Proton Conducting Materials Trivalent Metal Acid Sulphate Hydrates: Part I. The Vibrational Spectrum of H5O2+". In: *Journal of Molecular Structure* 195 (1989), pp. 283–291.

[178] E. S. Stoyanov and C. A. Reed. "IR Spectrum of the H5O2+ Cation in the Context of Proton Disolvates L−H+−L". In: *The Journal of Physical Chemistry A* 110.48 (2006), pp. 12992–13002.

[179] D. Decka, G. Schwaab, and M. Havenith. "A THz/FTIR Fingerprint of the Solvated Proton: Evidence for Eigen Structure and Zundel Dynamics". In: *Phys. Chem. Chem. Phys.* 17.17 (2015), pp. 11898–11907.

[180] M. Casula, C. Filippi, and S. Sorella. "Diffusion Monte Carlo Method with Lattice Regularization". In: *Phys. Rev. Lett.* 95.10 (2005), p. 100201.

[181] O. Markovitch, H. Chen, S. Izvekov, F. Paesani, G. A. Voth, and N. Agmon. "Special Pair Dance and Partner Selection: Elementary Steps in Proton Transport in Liquid Water". In: *The Journal of Physical Chemistry B* 112.31 (2008), pp. 9456–9466.

[182] K.-D. Kreuer, S. J. Paddison, E. Spohr, and M. Schuster. "Transport in Proton Conductors for Fuel-Cell Applications: Simulations, Elementary Reactions, and Phenomenology". In: *Chemical Reviews* 104.10 (2004), pp. 4637–4678.

[183] U. W. Schmitt and G. A. Voth. "The Computer Simulation of Proton Transport in Water". In: *The Journal of Chemical Physics* 111.20 (1999), pp. 9361–9381.

[184] R. M. Huchmala and A. B. McCoy. "Exploring the Origins of Spectral Signatures of Strong Hydrogen Bonding in Protonated Water Clusters". In: *The Journal of Physical Chemistry A* 126.8 (2022), pp. 1360–1368.

[185] P. B. Calio, C. Li, and G. A. Voth. "Resolving the Structural Debate for the Hydrated Excess Proton in Water". In: *Journal of the American Chemical Society* 143.44 (2021), pp. 18672–18683.

[186] J. O. Richardson and S. C. Althorpe. "Ring-Polymer Molecular Dynamics Rate-Theory in the Deep-Tunneling Regime: Connection with Semiclassical Instanton Theory". In: *The Journal of Chemical Physics* 131.21 (2009), p. 214106.

[187] A. I. Vaĭnshteĭn, V. I. Zakharov, V. A. Novikov, and M. A. Shifman. "ABC of Instantons". In: *Sov. Phys. Usp.* 25.4 (1982), pp. 195–215.

[188] Z. Jiang, V. N. Smelyanskiy, S. V. Isakov, S. Boixo, G. Mazzola, M. Troyer, and H. Neven. "Scaling Analysis and Instantons for Thermally Assisted Tunneling and Quantum Monte Carlo Simulations". In: *Phys. Rev. A* 95.1 (2017), p. 382.

[189] G. Mazzola, V. N. Smelyanskiy, and M. Troyer. "Quantum Monte Carlo Tunneling from Quantum Chemistry to Quantum Annealing". In: *Phys. Rev. B* 96.13 (2017), p. 483.

[190] T. J. Hele. "On the Relation between Thermostatted Ring-Polymer Molecular Dynamics and Exact Quantum Dynamics". In: *Molecular Physics* 114.9 (2016), pp. 1461–1471.

[191] H. Arcis, J. Plumridge, and P. R. Tremaine. "Limiting Conductivities of Strong Acids and Bases in D2O and H2O: Deuterium Isotope Effects on Proton Hopping over a Wide Temperature Range". In: *The Journal of Physical Chemistry B* 126.43 (2022), pp. 8791–8803.

[192] P. J. Linstorm and W. G. Mallard. "NIST Chemistry Webbook, NIST Standard Reference Database Number 69". In: ().

[193] D. Chandler and K. Leung. "Excess Electrons in Liquids: Geometrical Perspectives". In: *Annu. Rev. Phys. Chem.* 45.1 (1994), pp. 557–591.

[194] F. Dahms, B. P. Fingerhut, E. T. J. Nibbering, E. Pines, and T. Elsaesser. "Large-Amplitude Transfer Motion of Hydrated Excess Protons Mapped by Ultrafast 2D IR Spectroscopy". In: *Science* (2017).

[195] J. A. Fournier, W. B. Carpenter, N. H. C. Lewis, and A. Tokmakoff. "Broadband 2D IR Spectroscopy Reveals Dominant Asymmetric H5O2+ Proton Hydration Structures in Acid Solutions". In: *Nature Chemistry* 10.9 (2018), pp. 932–937.

[196] C. A. J. Daly, L. M. Streacker, Y. Sun, S. R. Pattenaude, A. A. Hassanali, P. B. Petersen, S. A. Corcelli, and D. Ben-Amotz. "Decomposition of the Experimental Raman and Infrared Spectra of Acidic Water into Proton, Special Pair, and Counterion Contributions". In: *The Journal of Physical Chemistry Letters* 8.21 (2017), pp. 5246–5252.

[197] W. B. Carpenter, Q. Yu, J. H. Hack, B. Dereka, J. M. Bowman, and A. Tokmakoff. "Decoding the 2D IR Spectrum of the Aqueous Proton with High-Level VSCF/VCI Calculations". In: *The Journal of Chemical Physics* 153.12 (2020), p. 124506.

[198] B. Dereka, Q. Yu, N. H. C. Lewis, W. B. Carpenter, J. M. Bowman, and A. Tokmakoff. "Crossover from Hydrogen to Chemical Bonding". In: *Science* 371.6525 (2021), pp. 160–164.

[199] H. Lapid, N. Agmon, M. K. Petersen, and G. A. Voth. "A Bond-Order Analysis of the Mechanism for Hydrated Proton Mobility in Liquid Water". In: *The Journal of Chemical Physics* 122.1 (2005), p. 014506.

[200] I. G. Kaplan. *Intermolecular Interactions: Physical Picture, Computational Methods and Model Potentials*. Vol. 128. Chichester: John Wiley & Sons, Ltd, 2006. 368 pp.

[201] J. F. Ouyang and R. P. A. Bettens. "Modelling Water: A Lifetime Enigma". In: *CHIMIA* 69.3 (3 2015), pp. 104–104.

[202] B. Guillot. "A Reappraisal of What We Have Learnt during Three Decades of Computer Simulations on Water". In: *Journal of Molecular Liquids*. Molecular Liquids. Water at the New Millenium 101.1 (2002), pp. 219–260.

[203] P. Mark and L. Nilsson. "Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K". In: *J. Phys. Chem. A* 105.43 (2001), pp. 9954–9960.

[204] S. P. Kadaoluwa Pathirannahalage, N. Meftahi, A. Elbourne, A. C. G. Weiss, C. F. McConville, A. Padua, D. A. Winkler, M. Costa Gomes, T. L. Greaves, T. C. Le, Q. A. Besford, and A. J. Christofferson. "Systematic Comparison of the Structural and Dynamic Properties of Commonly Used Water Models for Molecular Dynamics Simulations". In: *Journal of Chemical Information and Modeling* 61.9 (2021), pp. 4521–4536.

[205] T. Plé, L. Lagardère, and J.-P. Piquemal. "Force-Field-Enhanced Neural Network Interactions: From Local Equivariant Embedding to Atom-in-Molecule Properties and Long-Range Effects". In: *Chemical Science* 14.44 (2023), pp. 12554–12569.

[206] J. E. Lennard-Jones. "Cohesion". In: *Proceedings of the Physical Society* 43.5 (1931), p. 461.

[207] W. L. Jorgensen. "Quantum and Statistical Mechanical Studies of Liquids. 10. Transferable Intermolecular Potential Functions for Water, Alcohols, and Ethers. Application to Liquid Water". In: *Journal of the American Chemical Society* 103.2 (1981), pp. 335–340.

[208] W. L. Jorgensen and J. D. Madura. "Quantum and Statistical Mechanical Studies of Liquids. 25. Solvation and Conformation of Methanol in Water". In: *J. Am. Chem. Soc.* 105.6 (1983), pp. 1407–1413.

[209] M. W. Mahoney and W. L. Jorgensen. "A Five-Site Model for Liquid Water and the Reproduction of the Density Anomaly by Rigid, Nonpolarizable Potential Functions". In: *J. Chem. Phys.* 112.20 (2000), pp. 8910–8922.

[210] J. L. F. Abascal and C. Vega. "A General Purpose Model for the Condensed Phases of Water: TIP4P/2005". In: *The Journal of Chemical Physics* 123.23 (2005), p. 234505.

[211] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans. "Interaction Models for Water in Relation to Protein Hydration". In: *Intermolecular Forces: Proceedings of the Fourteenth Jerusalem Symposium on Quantum Chemistry and Biochemistry Held in Jerusalem, Israel, April 13–16, 1981*. Ed. by B. Pullman. Dordrecht: Springer Netherlands, 1981, pp. 331–342.

[212] M. W. Mahoney and W. L. Jorgensen. "Quantum, Intramolecular Flexibility, and Polarizability Effects on the Reproduction of the Density Anomaly of Liquid Water by Simple Potential Functions". In: *The Journal of Chemical Physics* 115.23 (2001), pp. 10758–10768.

[213] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma. "The Missing Term in Effective Pair Potentials". In: *The Journal of Physical Chemistry* 91.24 (1987), pp. 6269–6271.

[214] H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura, and T. Head-Gordon. "Development of an Improved Four-Site Water Model for Biomolecular Simulations: TIP4P-Ew". In: *The Journal of Chemical Physics* 120.20 (2004), pp. 9665–9678.

[215] S. Habershon, T. E. Markland, and D. E. Manolopoulos. "Competing Quantum Effects in the Dynamics of a Flexible Water Model". In: *The Journal of Chemical Physics* 131.2 (2009), p. 024501.

[216] F. Paesani, W. Zhang, D. A. Case, T. E. Cheatham III, and G. A. Voth. "An Accurate and Simple Quantum Model for Liquid Water". In: *The Journal of Chemical Physics* 125.18 (2006), p. 184507.

[217] A. Stone. *The Theory of Intermolecular Forces*. Oxford University Press, 2013.

[218] C. Millot and A. J. Stone. "Towards an Accurate Intermolecular Potential for Water". In: *Molecular Physics* 77.3 (1992), pp. 439–462.

[219] E. M. Mas, K. Szalewicz, R. Bukowski, and B. Jeziorski. "Pair Potential for Water from Symmetry-Adapted Perturbation Theory". In: *The Journal of Chemical Physics* 107.11 (1997), pp. 4207–4218.

[220] C. J. Burnham, J. Li, S. S. Xantheas, and M. Leslie. "The Parametrization of a Thole-type All-Atom Polarizable Water Model from First Principles and Its Application to the Study of Water Clusters (N=2–21) and the Phonon Spectrum of Ice Ih". In: *The Journal of Chemical Physics* 110.9 (1999), pp. 4566–4581.

[221] C. J. Burnham and S. S. Xantheas. "Development of Transferable Interaction Models for Water. IV. A Flexible, All-Atom Polarizable Potential (TTM2-F) Based on Geometry Dependent Charges Derived from an Ab Initio Monomer Dipole Moment Surface". In: *The Journal of Chemical Physics* 116.12 (2002), pp. 5115–5124.

[222] S. S. Xantheas, C. J. Burnham, and R. J. Harrison. "Development of Transferable Interaction Models for Water. II. Accurate Energetics of the First Few Water Clusters from First Principles". In: *The Journal of Chemical Physics* 116.4 (2002), pp. 1493–1499.

[223] G. S. Fanourgakis and S. S. Xantheas. "Development of Transferable Interaction Potentials for Water. V. Extension of the Flexible, Polarizable, Thole-type Model Potential (TTM3-F, v. 3.0) to Describe the Vibrational Spectra of Water Clusters and Liquid Water". In: *The Journal of Chemical Physics* 128.7 (2008), p. 074506.

[224] C. J. Burnham, D. J. Anick, P. K. Mankoo, and G. F. Reiter. "The Vibrational Proton Potential in Bulk Liquid Water and Ice". In: *The Journal of Chemical Physics* 128.15 (2008), p. 154519.

[225] R. K. Nesbet. "Atomic Bethe-Goldstone Equations". In: *Advances in Chemical Physics*. John Wiley & Sons, Ltd, 1969, pp. 1–34.

[226] R. Bukowski, K. Szalewicz, G. C. Groenenboom, and A. van der Avoird. "Predictions of the Properties of Water from First Principles". In: *Science* 315.5816 (2007), pp. 1249–1252.

[227] R. Bukowski, K. Szalewicz, G. C. Groenenboom, and A. van der Avoird. "Polarizable Interaction Potential for Water from Coupled Cluster Calculations. II. Applications to Dimer Spectra, Virial Coefficients, and Simulations of Liquid Water". In: *The Journal of Chemical Physics* 128.9 (2008), p. 094314.

[228] R. Bukowski, K. Szalewicz, G. C. Groenenboom, and A. van der Avoird. "Polarizable Interaction Potential for Water from Coupled Cluster Calculations. I. Analysis of Dimer Potential Energy Surface". In: *The Journal of Chemical Physics* 128.9 (2008), p. 094313.

[229] R. Kumar, F.-F. Wang, G. R. Jenness, and K. D. Jordan. "A Second Generation Distributed Point Polarizable Water Model". In: *The Journal of Chemical Physics* 132.1 (2010), p. 014309.

[230] X. Huang, B. J. Braams, and J. M. Bowman. "Ab Initio Potential Energy and Dipole Moment Surfaces of (H2O)2". In: *The Journal of Physical Chemistry A* 110.2 (2006), pp. 445–451.

[231] X. Huang, B. J. Braams, J. M. Bowman, R. E. A. Kelly, J. Tennyson, G. C. Groenenboom, and A. van der Avoird. "New Ab Initio Potential Energy Surface and the Vibration-Rotation-Tunneling Levels of (H2O)2 and (D2O)2". In: *The Journal of Chemical Physics* 128.3 (2008), p. 034312.

[232] A. Shank, Y. Wang, A. Kaledin, B. J. Braams, and J. M. Bowman. "Accurate Ab Initio and "Hybrid" Potential Energy Surfaces, Intramolecular Vibrational Energies, and Classical Ir Spectrum of the Water Dimer". In: *The Journal of Chemical Physics* 130.14 (2009), p. 144314.

[233] V. Babin, C. Leforestier, and F. Paesani. "Development of a "First Principles" Water Potential with Flexible Monomers: Dimer Potential Energy Surface, VRT Spectrum, and Second Virial Coefficient". In: *J. Chem. Theory Comput.* 9.12 (2013), pp. 5395–5403.

[234] G. R. Medders, V. Babin, and F. Paesani. "A Critical Assessment of Two-Body and Three-Body Interactions in Water". In: *Journal of Chemical Theory and Computation* 9.2 (2013), pp. 1103–1114.

[235] V. Babin, G. R. Medders, and F. Paesani. "Development of a "First Principles" Water Potential with Flexible Monomers. II: Trimer Potential Energy Surface, Third Virial Coefficient, and Small Clusters". In: *J. Chem. Theory Comput.* 10.4 (2014), pp. 1599–1607.

[236] G. R. Medders, V. Babin, and F. Paesani. "Development of a "First-Principles" Water Potential with Flexible Monomers. III. Liquid Phase Properties". In: *J. Chem. Theory Comput.* 10.8 (2014), pp. 2906–2910.

[237] X. Zhu, M. Riera, E. F. Bull-Vulpe, and F. Paesani. "MB-pol(2023): Sub-chemical Accuracy for Water Simulations from the Gas to the Liquid Phase". In: *Journal of Chemical Theory and Computation* (2023).

[238] E. Palos, E. F. Bull-Vulpe, X. Zhu, H. Agnew, S. Gupta, and F. Paesani. *Current Status of the MB-pol Data-Driven Many-Body Potential for Predictive Simulations of Water Across Different Phases*. 2024. URL: `https://chemrxiv.org/engage/chemrxiv/article-details/66b9a345c9c6a5c07af7e70f`. Pre-published.

[239] A. Brown, B. J. Braams, K. Christoffel, Z. Jin, and J. M. Bowman. "Classical and Quasiclassical Spectral Analysis of CH5+ Using an Ab Initio Potential Energy Surface". In: *The Journal of Chemical Physics* 119.17 (2003), pp. 8790–8793.

[240] B. J. Braams and J. M. Bowman. "Permutationally Invariant Potential Energy Surfaces in High Dimensionality". In: *International Reviews in Physical Chemistry* 28.4 (2009), pp. 577–606.

[241] C. Qu, Q. Yu, and J. M. Bowman. "Permutationally Invariant Potential Energy Surfaces". In: *Annual Review of Physical Chemistry* 69.1 (2018), pp. 151–175.

[242] R. Conte, C. Qu, P. L. Houston, and J. M. Bowman. "Efficient Generation of Permutationally Invariant Potential Energy Surfaces for Large Molecules". In: *Journal of Chemical Theory and Computation* 16.5 (2020), pp. 3264–3272.

[243] T. T. Nguyen, E. Székely, G. Imbalzano, J. Behler, G. Csányi, M. Ceriotti, A. W. Götz, and F. Paesani. "Comparison of Permutationally Invariant Polynomials, Neural Networks, and Gaussian Approximation Potentials in Representing Water Interactions through Many-Body Expansions". In: *The Journal of Chemical Physics* 148.24 (2018), p. 241725.

[244] Q. Yu, C. Qu, P. L. Houston, A. Nandi, P. Pandey, R. Conte, and J. M. Bowman. "A Status Report on "Gold Standard" Machine-Learned Potentials for Water". In: *The Journal of Physical Chemistry Letters* 14.36 (2023), pp. 8077–8087.

[245] S. L. Bore and F. Paesani. "Realistic Phase Diagram of Water from "First Principles" Data-Driven Quantum Simulations". In: *Nature Communications* 14.1 (1 2023), p. 3349.

[246] E. Palos, S. Dasgupta, E. Lambros, and F. Paesani. "Data-Driven Many-Body Potentials from Density Functional Theory for Aqueous Phase Chemistry". In: *Chemical Physics Reviews* 4.1 (2023), p. 011301.

[247] Y. Zhai, A. Caruso, S. L. Bore, Z. Luo, and F. Paesani. "A "Short Blanket" Dilemma for a State-of-the-Art Neural Network Potential for Water: Reproducing Experimental Properties or the Physics of the Underlying Many-Body Interactions?" In: *The Journal of Chemical Physics* 158.8 (2023), p. 084111.

[248] G. A. Cisneros, K. T. Wikfeldt, L. Ojamäe, J. Lu, Y. Xu, H. Torabifard, A. P. Bartók, G. Csányi, V. Molinero, and F. Paesani. "Modeling Molecular Interactions in Water: From Pairwise to Many-Body Potential Energy Functions". In: *Chemical Reviews* 116.13 (2016), pp. 7501–7528.

[249] Y. Wang, X. Huang, B. C. Shepler, B. J. Braams, and J. M. Bowman. "Flexible, Ab Initio Potential, and Dipole Moment Surfaces for Water. I. Tests and Applications for Clusters up to the 22-Mer". In: *The Journal of Chemical Physics* 134.9 (2011), p. 094509.

[250] H. Partridge and D. W. Schwenke. "The Determination of an Accurate Isotope Dependent Potential Energy Surface for Water from Extensive Ab Initio Calculations and Experimental Data". In: *The Journal of Chemical Physics* 106.11 (1997), pp. 4618–4639.

[251] Q. Yu, C. Qu, P. L. Houston, R. Conte, A. Nandi, and J. M. Bowman. "Q-AQUA: A Many-Body CCSD(T) Water Potential, Including Four-Body Interactions, Demonstrates the Quantum Nature of Water from Clusters to the Liquid Phase". In: *The Journal of Physical Chemistry Letters* 13.22 (2022), pp. 5068–5074.

[252] C. Qu, Q. Yu, P. L. Houston, R. Conte, A. Nandi, and J. M. Bowman. "Interfacing Q-AQUA with a Polarizable Force Field: The Best of Both Worlds". In: *Journal of Chemical Theory and Computation* 19.12 (2023), pp. 3446–3459.

[253] C. K. Egan and F. Paesani. "Assessing Many-Body Effects of Water Self-Ions. I: OH–(H2O)n Clusters". In: *Journal of Chemical Theory and Computation* 14.4 (2018), pp. 1982–1997.

[254] C. K. Egan and F. Paesani. "Assessing Many-Body Effects of Water Self-Ions. II: H3O+(H2O)n Clusters". In: *Journal of Chemical Theory and Computation* 15.9 (2019), pp. 4816–4833.

[255] X. Huang, B. J. Braams, and J. M. Bowman. "Ab Initio Potential Energy and Dipole Moment Surfaces for H5O2+". In: *J. Chem. Phys.* 122.4 (2005), p. 044308.

[256] Q. Yu and J. M. Bowman. "Ab Initio Potential for H3O+ → H+ + H2O: A Step to a Many-Body Representation of the Hydrated Proton?" In: *Journal of Chemical Theory and Computation* 12.11 (2016), pp. 5284–5292.

[257] Q. Yu and J. M. Bowman. "Tracking Hydronium/Water Stretches in Magic H3O+(H2O)20 Clusters through High-level Quantum VSCF/VCI Calculations". In: *The Journal of Physical Chemistry A* 124.6 (2020), pp. 1167–1175.

[258] Q. Yu and J. M. Bowman. "How the Zundel (H5O2+) Potential Can Be Used to Predict the Proton Stretch and Bend Frequencies of Larger Protonated Water Clusters". In: *The Journal of Physical Chemistry Letters* 7.24 (2016), pp. 5259–5265.

[259] Q. Yu and J. M. Bowman. "Communication: VSCF/VCI Vibrational Spectroscopy of H7 O 3 + and H9 O 4 + Using High-Level, Many-Body Potential Energy Surface and Dipole Moment Surfaces". In: *The Journal of Chemical Physics* 146.12 (2017), p. 121102.

[260] Q. Yu, W. B. Carpenter, N. H. C. Lewis, A. Tokmakoff, and J. M. Bowman. "High-Level VSCF/VCI Calculations Decode the Vibrational Spectrum of the Aqueous Proton". In: *The Journal of Physical Chemistry B* 123.33 (2019), pp. 7214–7224.

[261] P. Pinski and G. Csányi. "Reactive Many-Body Expansion for a Protonated Water Cluster". In: *Journal of Chemical Theory and Computation* 10.1 (2014), pp. 68–75.

[262] W. Zhang and A. C. T. van Duin. "Second-Generation ReaxFF Water Force Field: Improvements in the Description of Water Density and OH-Anion Diffusion". In: *The Journal of Physical Chemistry B* 121.24 (2017), pp. 6021–6032.

[263] J. Lobaugh and G. A. Voth. "The Quantum Dynamics of an Excess Proton in Water". In: *The Journal of Chemical Physics* 104.5 (1996), pp. 2056–2069.

[264] U. W. Schmitt and G. A. Voth. "Multistate Empirical Valence Bond Model for Proton Transport in Water". In: *The Journal of Physical Chemistry B* 102.29 (1998), pp. 5547–5551.

[265] T. J. F. Day, A. V. Soudackov, M. Čuma, U. W. Schmitt, and G. A. Voth. "A Second Generation Multistate Empirical Valence Bond Model for Proton Transport in Aqueous Systems". In: *The Journal of Chemical Physics* 117.12 (2002), pp. 5839–5849.

[266] G. A. Voth. "Computer Simulation of Proton Solvation and Transport in Aqueous and Biomolecular Systems". In: *Accounts of Chemical Research* 39.2 (2006), pp. 143–150.

[267] R. Biswas, W. Carpenter, J. A. Fournier, G. A. Voth, and A. Tokmakoff. "IR Spectral Assignments for the Hydrated Excess Proton in Liquid Water". In: *The Journal of Chemical Physics* 146.15 (2017), p. 154507.

[268] W. Pronobis and K.-R. Müller. "Kernel Methods for Quantum Chemistry". In: *Machine Learning Meets Quantum Physics*. Ed. by K. T. Schütt, S. Chmiela, O. A. von Lilienfeld, A. Tkatchenko, K. Tsuda, and K.-R. Müller. Cham: Springer International Publishing, 2020, pp. 25–36.

[269] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld. "Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning". In: *Physical Review Letters* 108.5 (2012), p. 058301.

[270] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, and K.-R. Müller. "Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies". In: *Journal of Chemical Theory and Computation* 9.8 (2013), pp. 3404–3419.

[271] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko. "Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space". In: *The Journal of Physical Chemistry Letters* 6.12 (2015), pp. 2326–2331.

[272] R. Drautz, M. Fähnle, and J. M. Sanchez. "General Relations between Many-Body Potentials and Cluster Expansions in Multicomponent Systems". In: *Journal of Physics: Condensed Matter* 16.23 (2004), p. 3843.

[273] E. Prodan and W. Kohn. "Nearsightedness of Electronic Matter". In: *Proceedings of the National Academy of Sciences* 102.33 (2005), pp. 11635–11638.

[274] D. M. Anstine and O. Isayev. "Machine Learning Interatomic Potentials and Long-Range Physics". In: *The Journal of Physical Chemistry A* 127.11 (2023), pp. 2417–2431.

[275] J. Behler and M. Parrinello. "Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces". In: *Physical Review Letters* 98.14 (2007), p. 146401.

[276] J. Behler. "Atom-Centered Symmetry Functions for Constructing High-Dimensional Neural Network Potentials". In: *The Journal of Chemical Physics* 134.7 (2011), p. 074106.

[277] A. P. Bartók, R. Kondor, and G. Csányi. "On Representing Chemical Environments". In: *Physical Review B* 87.18 (2013), p. 184115.

[278] F. A. Faber, A. S. Christensen, B. Huang, and O. A. von Lilienfeld. "Alchemical and Structural Distribution Based Representation for Universal Quantum Machine Learning". In: *The Journal of Chemical Physics* 148.24 (2018), p. 241717.

[279] A. S. Christensen, L. A. Bratholm, F. A. Faber, and O. Anatole von Lilienfeld. "FCHL Revisited: Faster and More Accurate Quantum Machine Learning". In: *The Journal of Chemical Physics* 152.4 (2020), p. 044107.

[280] S. N. Pozdnyakov, M. J. Willatt, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti. "Incompleteness of Atomic Structure Representations". In: *Physical Review Letters* 125.16 (2020), p. 166001.

[281] A. V. Shapeev. "Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials". In: *Multiscale Modeling & Simulation* 14.3 (2016), pp. 1153–1173.

[282] R. Drautz. "Atomic Cluster Expansion for Accurate and Transferable Interatomic Potentials". In: *Physical Review B* 99.1 (2019), p. 014104.

[283] Y. Lysogorskiy, C. van der Oord, A. Bochkarev, S. Menon, M. Rinaldi, T. Hammerschmidt, M. Mrovec, A. Thompson, G. Csányi, C. Ortner, and R. Drautz. "Performant Implementation of the Atomic Cluster Expansion (PACE) and Application to Copper and Silicon". In: *npj Computational Materials* 7.1 (1 2021), pp. 1–12.

[284] J. Nigam, S. Pozdnyakov, G. Fraux, and M. Ceriotti. "Unified Theory of Atom-Centered Representations and Message-Passing Machine-Learning Schemes". In: *The Journal of Chemical Physics* 156.20 (2022), p. 204115.

[285] M. F. Langer, A. Goeßmann, and M. Rupp. "Representations of Molecules and Materials for Interpolation of Quantum-Mechanical Simulations via Machine Learning". 2021.

[286] F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti. "Physics-Inspired Structural Representations for Molecules and Materials". In: *Chemical Reviews* 121.16 (2021), pp. 9759–9815.

[287] M. Uhrin. "Through the Eyes of a Descriptor: Constructing Complete, Invertible Descriptions of Atomic Environments". In: *Physical Review B* 104.14 (2021), p. 144110.

[288] M. J. Willatt, F. Musil, and M. Ceriotti. "Atom-Density Representations for Machine Learning". In: *The Journal of Chemical Physics* 150.15 (2019), p. 154110.

[289] V. N. Vapnik. *The Nature of Statistical Learning Theory*. New York, NY: Springer, 2000.

[290] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge: Cambridge University Press, 2014.

[291] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2018.

[292] G. Wahba. *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, 1990. 177 pp.

[293] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. Red. by F. Bach. Adaptive Computation and Machine Learning Series. Cambridge, MA, USA: MIT Press, 2005. 272 pp.

[294] A. P. Bartók and G. Csányi. "Gaussian Approximation Potentials : A Brief Tutorial Introduction". In: *International Journal of Quantum Chemistry* 115.16 (16 2015), pp. 1051–1057.

[295] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller. "Machine Learning of Accurate Energy-Conserving Molecular Force Fields". In: *Science Advances* 3.5 (2017), e1603015.

[296] A. S. Christensen, F. A. Faber, and O. A. von Lilienfeld. "Operators in Quantum Machine Learning: Response Properties in Chemical Space". In: *The Journal of Chemical Physics* 150.6 (2019), p. 064105.

[297] O. A. von Lilienfeld. "Quantum Machine Learning in Chemical Compound Space". In: *Angewandte Chemie International Edition* 57.16 (2018), pp. 4164–4169.

[298] W. S. McCulloch and W. Pitts. "A Logical Calculus of the Ideas Immanent in Nervous Activity". In: *The bulletin of mathematical biophysics* 5.4 (1943), pp. 115–133.

[299] F. Rosenblatt. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain". In: *Psychological Review* 65.6 (1958), pp. 386–408.

[300] M. Minsky and S. A. Papert. *Perceptrons: An Introduction to Computational Geometry*. The MIT Press, 2017.

[301] BrunelloN. *Example of a Neuron https://commons.wikimedia.org/wiki/File:Example_of_a_neuron.png*. 2021.

[302] L. Bottou. "Large-Scale Machine Learning with Stochastic Gradient Descent". In: *Proceedings of COMPSTAT'2010*. Ed. by Y. Lechevallier and G. Saporta. Heidelberg: Physica-Verlag HD, 2010, pp. 177–186.

[303] C. M. Bishop. "Training with Noise Is Equivalent to Tikhonov Regularization". In: *Neural Computation* 7.1 (1995), pp. 108–116.

[304] D. P. Kingma and J. Ba. *Adam: A Method for Stochastic Optimization*. 2017. URL: `http://arxiv.org/abs/1412.6980`. Pre-published.

[305] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. "Learning Representations by Back-Propagating Errors". In: *Nature* 323.6088 (6088 1986), pp. 533–536.

[306] T. B. Blank, S. D. Brown, A. W. Calhoun, and D. J. Doren. "Neural Network Models of Potential Energy Surfaces". In: *The Journal of Chemical Physics* 103.10 (1995), pp. 4129–4137.

[307] H. Gassner, M. Probst, A. Lauenstein, and K. Hermansson. "Representation of Intermolecular Potential Functions by Neural Networks". In: *The Journal of Physical Chemistry A* 102.24 (1998), pp. 4596–4605.

[308] S. Hobday, R. Smith, and J. Belbruno. "Applications of Neural Networks to Fitting Interatomic Potential Functions". In: *Modelling and Simulation in Materials Science and Engineering* 7.3 (1999), p. 397.

[309] S. Manzhos and T. Carrington Jr. "Using Redundant Coordinates to Represent Potential Energy Surfaces with Lower-Dimensional Functions". In: *The Journal of Chemical Physics* 127.1 (2007), p. 014103.

[310] M. Malshe, R. Narulkar, L. M. Raff, M. Hagan, S. Bukkapatnam, P. M. Agrawal, and R. Komanduri. "Development of Generalized Potential-Energy Surfaces Using Many-Body Expansions, Neural Networks, and Moiety Energy Approximations". In: *The Journal of Chemical Physics* 130.18 (2009), p. 184102.

[311] B. Jiang, J. Li, and H. Guo. "Potential Energy Surfaces from High Fidelity Fitting of Ab Initio Points: The Permutation Invariant Polynomial - Neural Network Approach". In: *International Reviews in Physical Chemistry* 35.3 (2016), pp. 479–506.

[312] J. Behler. "Four Generations of High-Dimensional Neural Network Potentials". In: *Chemical Reviews* 121.16 (2021), pp. 10037–10072.

[313] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. "The Limitations of Deep Learning in Adversarial Settings". In: *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*. 2016 IEEE European Symposium on Security and Privacy (EuroS&P). 2016, pp. 372–387.

[314] J. Behler. "Constructing High-Dimensional Neural Network Potentials: A Tutorial Review". In: *International Journal of Quantum Chemistry* 115.16 (2015), pp. 1032–1050.

[315] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. "Neural Message Passing for Quantum Chemistry". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. Sydney, NSW, Australia: JMLR.org, 2017, pp. 1263–1272.

[316] W. L. Hamilton. *Graph Representation Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Cham: Springer International Publishing, 2020.

[317] P. Reiser, M. Neubert, A. Eberhard, L. Torresi, C. Zhou, C. Shao, H. Metni, C. van Hoesel, H. Schopmans, T. Sommer, and P. Friederich. "Graph Neural Networks for Materials Science and Chemistry". In: *Communications Materials* 3.1 (2022), pp. 1–18.

[318] I. Batatia, D. P. Kovacs, G. Simm, C. Ortner, and G. Csanyi. "MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 11423–11436.

[319] A. Omranpour, P. M. De Hijes, J. Behler, and C. Dellago. *Perspective: Atomistic Simulations of Water and Aqueous Systems with Machine Learning Potentials*. 2024. URL: http://arxiv.org/abs/2401.17875. Pre-published.

[320] P. Montero de Hijes, C. Dellago, R. Jinnouchi, B. Schmiedmayer, and G. Kresse. "Comparing Machine Learning Potentials for Water: Kernel-based Regression and Behler–Parrinello Neural Networks". In: *The Journal of Chemical Physics* 160.11 (2024), p. 114107.

[321] T. Morawietz, A. Singraber, C. Dellago, and J. Behler. "How van Der Waals Interactions Determine the Unique Properties of Water". In: *Proceedings of the National Academy of Sciences* 113.30 (2016), pp. 8368–8373.

[322] I. Batatia, P. Benner, Y. Chiang, A. M. Elena, D. P. Kovács, J. Riebesell, X. R. Advincula, M. Asta, M. Avaylon, W. J. Baldwin, F. Berger, N. Bernstein, A. Bhowmik, S. M. Blau, V. Cărare, J. P. Darby, S. De, F. Della Pia, V. L. Deringer, R. Elijošius, Z. El-Machachi, F. Falcioni, E. Fako, A. C. Ferrari, A. Genreith-Schriever, J. George, R. E. A. Goodall, C. P. Grey, P. Grigorev, S. Han, W. Handley, H. H. Heenen, K. Hermansson, C. Holm, J. Jaafar, S. Hofmann, K. S. Jakob, H. Jung, V. Kapil, A. D. Kaplan, N. Karimitari, J. R. Kermode, N. Kroupa, J. Kullgren, M. C. Kuner, D. Kuryla, G. Liepuoniute, J. T. Margraf, I.-B. Magdău, A. Michaelides, J. H. Moore, A. A. Naik, S. P. Niblett, S. W. Norwood, N. O'Neill, C.

Ortner, K. A. Persson, K. Reuter, A. S. Rosen, L. L. Schaaf, C. Schran, B. X. Shi, E. Sivonxay, T. K. Stenczel, V. Svahn, C. Sutton, T. D. Swinburne, J. Tilly, C. van der Oord, E. Varga-Umbrich, T. Vegge, M. Vondrák, Y. Wang, W. C. Witt, F. Zills, and G. Csányi. *A Foundation Model for Atomistic Materials Chemistry*. 2024. URL: `http://arxiv.org/abs/2401.00096`. Pre-published.

[323] L. Zhang, H. Wang, R. Car, and E. Weinan. "Phase Diagram of a Deep Potential Water Model". In: *Physical review letters* 126.23 (2021), p. 236001.

[324] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld. "Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach". In: *Journal of Chemical Theory and Computation* 11.5 (2015), pp. 2087–2096.

[325] M. S. Chen, J. Lee, H.-Z. Ye, T. C. Berkelbach, D. R. Reichman, and T. E. Markland. "Data-Efficient Machine Learning Potentials from Transfer Learning of Periodic Correlated Electronic Structure Methods: Liquid Water at AFQMC, CCSD, and CCSD(T) Accuracy". In: *Journal of Chemical Theory and Computation* 19.14 (2023), pp. 4510–4519.

[326] J. Öström. "A Flexible and Polarizable Water Model Built on Interpolated Multipoles". Department of Physics, Stockholm University, 2023.

[327] J. Daru, H. Forbert, J. Behler, and D. Marx. "Coupled Cluster Molecular Dynamics of Condensed Phase Systems Enabled by Machine Learning Potentials: Liquid Water Benchmark". In: *Physical Review Letters* 129.22 (2022), p. 226001.

[328] S. Yue, M. C. Muniz, M. F. Calegari Andrade, L. Zhang, R. Car, and A. Z. Panagiotopoulos. "When Do Short-Range Atomistic Machine-Learning Models Fall Short?" In: *The Journal of Chemical Physics* 154.3 (2021), p. 034111.

[329] A. Grisafi and M. Ceriotti. "Incorporating Long-Range Physics in Atomic-Scale Machine Learning". In: *The Journal of Chemical Physics* 151.20 (2019), p. 204105.

[330] T. W. Ko, J. A. Finkler, S. Goedecker, and J. Behler. "Accurate Fourth-Generation Machine Learning Potentials by Electrostatic Embedding". In: *Journal of Chemical Theory and Computation* 19.12 (2023), pp. 3567–3579.

[331] A. Gao and R. C. Remsing. "Self-Consistent Determination of Long-Range Electrostatics in Neural Network Potentials". In: *Nature Communications* 13.1 (1 2022), p. 1572.

[332] H. S. Dhattarwal, A. Gao, and R. C. Remsing. "Dielectric Saturation in Water from a Long-Range Machine Learning Model". In: *The Journal of Physical Chemistry B* 127.16 (2023), pp. 3663–3671.

[333] L. Zhang, M. Chen, X. Wu, H. Wang, W. E, and R. Car. "Deep Neural Network for the Dielectric Response of Insulators". In: *Physical Review B* 102.4 (2020), p. 041121.

[334] L. Zhang, H. Wang, M. C. Muniz, A. Z. Panagiotopoulos, R. Car, and W. E. "A Deep Potential Model with Long-Range Electrostatic Interactions". In: *The Journal of Chemical Physics* 156.12 (2022).

[335] M. Calegari Andrade, R. Car, and A. Selloni. "Probing the Self-Ionization of Liquid Water with Ab Initio Deep Potential Molecular Dynamics". In: *Proceedings of the National Academy of Sciences* 120.46 (2023), e2302468120.

[336] S. Dasgupta, G. Cassone, and F. Paesani. *Nuclear Quantum Effects and the Grotthuss Mechanism Dictate the pH of Liquid Water*. 2024. URL: https://chemrxiv.org/engage/chemrxiv/article-details/6660bc25418a5379b02924d5. Pre-published.

[337] S. K. Natarajan, T. Morawietz, and J. Behler. "Representing the Potential-Energy Surface of Protonated Water Clusters by High-Dimensional Neural Network Potentials". In: *Physical Chemistry Chemical Physics* 17.13 (2015), pp. 8356–8371.

[338] C. Schran, F. Brieuc, and D. Marx. "Converged Colored Noise Path Integral Molecular Dynamics Study of the Zundel Cation Down to Ultralow Temperatures at Coupled Cluster Accuracy". In: *J. Chem. Theory Comput.* 14.10 (2018), pp. 5068–5078.

[339] C. Schran, F. Uhl, J. Behler, and D. Marx. "High-Dimensional Neural Network Potentials for Solvation: The Case of Protonated Water Clusters in Helium". In: *The Journal of Chemical Physics* 148.10 (2018), p. 102310.

[340] C. Schran and D. Marx. "Quantum Nature of the Hydrogen Bond from Ambient Conditions down to Ultra-Low Temperatures". In: *Physical Chemistry Chemical Physics* 21.45 (2019), pp. 24967–24975.

[341] C. Schran, J. Behler, and D. Marx. "Automated Fitting of Neural Network Potentials at Coupled Cluster Accuracy: Protonated Water Clusters as Testing Ground". In: *Journal of Chemical Theory and Computation* 16.1 (2020), pp. 88–99.

[342] C. Schran, F. Brieuc, and D. Marx. "Transferability of Machine Learning Potentials: Protonated Water Neural Network Potential Applied to the Protonated Water Hexamer". In: *The Journal of Chemical Physics* 154.5 (2021), p. 051101.

[343] H. Niu, Y. Yang, S. Jensen, M. Holzmann, C. Pierleoni, and D. M. Ceperley. "Stable Solid Molecular Hydrogen above 900 K from a Machine-Learned Potential Trained with Diffusion Quantum Monte Carlo". In: *Phys. Rev. Lett.* 130.7 (2023), p. 076102.

[344] A. Tirelli, G. Tenti, K. Nakano, and S. Sorella. "High-Pressure Hydrogen by Machine Learning and Quantum Monte Carlo". In: *Physical Review B* 106.4 (2022), p. L041105.

[345] G. Tenti, A. Tirelli, K. Nakano, M. Casula, and S. Sorella. *Principal Deuterium Hugoniot via Quantum Monte Carlo and Δ-learning*. 2023. URL: http://arxiv.org/abs/2301.03570. Pre-published.

[346] B. Huang, O. A. von Lilienfeld, J. T. Krogel, and A. Benali. "Toward DMC Accuracy Across Chemical Space with Scalable Δ-QML". In: *Journal of Chemical Theory and Computation* 19.6 (2023), pp. 1711–1721.

[347] C. Huang and B. M. Rubenstein. "Machine Learning Diffusion Monte Carlo Forces". In: *The Journal of Physical Chemistry A* 127.1 (2023), pp. 339–355.

[348] D. M. Ceperley, S. Jensen, Y. Yang, H. Niu, C. Pierleoni, and M. Holzmann. "Training Models Using Forces Computed by Stochastic Electronic Structure Methods". In: *Electronic Structure* 6.1 (2024), p. 015011.

[349] E. Slootman, I. Poltavsky, R. Shinde, J. Cocomello, S. Moroni, A. Tkatchenko, and C. Filippi. "Accurate Quantum Monte Carlo Forces for Machine-Learned Force Fields: Ethanol as a Benchmark". In: *Journal of Chemical Theory and Computation* (2024).

[350] X. Fu, Z. Wu, W. Wang, T. Xie, S. Keten, R. Gomez-Bombarelli, and T. Jaakkola. *Forces Are Not Enough: Benchmark and Critical Evaluation for Machine Learning Force Fields with Molecular Simulations*. 2023. URL: http://arxiv.org/abs/2210.07237. Pre-published.

[351] J. D. Morrow, J. L. A. Gardner, and V. L. Deringer. "How to Validate Machine-Learned Interatomic Potentials". In: *The Journal of Chemical Physics* 158.12 (2023), p. 121501.

[352] D. Wei and D. R. Salahub. "Hydrated Proton Clusters and Solvent Effects on the Proton Transfer Barrier: A Density Functional Study". In: *The Journal of Chemical Physics* 101.9 (1994), pp. 7633–7642.

[353] M. Park, I. Shin, N. J. Singh, and K. S. Kim. "Eigen and Zundel Forms of Small Protonated Water Clusters: Structures and Infrared Spectra". In: *The Journal of Physical Chemistry A* 111.42 (2007), pp. 10692–10702.

[354] D. J. Wales. "Rearrangements and Tunneling Splittings of Protonated Water Dimer". In: *J. Chem. Phys.* 110.21 (1999), pp. 10403–10409.

[355] A. A. Auer, T. Helgaker, and W. Klopper. "Accurate Molecular Geometries of the Protonated Water Dimer". In: *Phys. Chem. Chem. Phys.* 2.10 (2000), pp. 2235–2238.

[356] E. F. Valeev and H. F. Schaefer III. "The Protonated Water Dimer: Brueckner Methods Remove the Spurious C1 Symmetry Minimum". In: *The Journal of Chemical Physics* 108.17 (1998), pp. 7197–7201.

[357] Y. Xie, R. B. Remington, and H. F. Schaefer. "The Protonated Water Dimer: Extensive Theoretical Studies of H5O+2". In: *J. Chem. Phys.* 101.6 (1994), pp. 4878–4884.

[358] V. Kapil, J. VandeVondele, and M. Ceriotti. "Accurate Molecular Dynamics and Nuclear Quantum Effects at Low Cost by Multiple Steps in Real and Imaginary Time: Using Density Functional Theory to Accelerate Wavefunction Methods". In: *J. Chem. Phys.* 144.5 (2016), p. 054111.

[359] J. M. Headrick, E. G. Diken, R. S. Walters, N. I. Hammer, R. A. Christie, J. Cui, E. M. Myshakin, M. A. Duncan, M. A. Johnson, and K. D. Jordan. "Spectral Signatures of Hydrated Proton Vibrations in Water Clusters". In: *Science* 308.5729 (2005), pp. 1765–1769.

[360] R. Vuilleumier and D. Borgis. "Molecular Dynamics of an Excess Proton in Water Using a Non-Additive Valence Bond Force Field". In: *Journal of Molecular Structure*. Structure, Properties and Dynamics of Molecular Systems 436–437 (1997), pp. 555–565.

[361] R. Vuilleumier and D. Borgis. "An Extended Empirical Valence Bond Model for Describing Proton Transfer in H+(H2O)n Clusters and Liquid Water". In: *Chemical Physics Letters* 284.1 (1998), pp. 71–77.

[362] R. Vuilleumier and D. Borgis. "Wavefunction Quantization of the Proton Motion in a H5O2+ Dimer Solvated in Liquid Water". In: *Journal of molecular structure* 552.1/3 (2000), pp. 117–136.

[363] B. A. Helfrecht, R. K. Cersonsky, G. Fraux, and M. Ceriotti. "Structure-Property Maps with Kernel Principal Covariates Regression". In: *Machine Learning: Science and Technology* 1.4 (2020), p. 045021.

[364] F. Musil, M. Veit, A. Goscinski, G. Fraux, M. J. Willatt, M. Stricker, T. Junge, and M. Ceriotti. "Efficient Implementation of Atom-Density Representations". In: *The Journal of Chemical Physics* 154.11 (2021), p. 114109.

[365] L. Hascoet and V. Pascual. "The Tapenade Automatic Differentiation Tool: Principles, Model, and Specification". In: *ACM Transactions on Mathematical Software* 39.3 (2013), 20:1–20:43.

[366] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. "Scikit-Learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.85 (2011), pp. 2825–2830.

[367] G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, and M. Ceriotti. "Automatic Selection of Atomic Fingerprints and Reference Configurations for Machine-Learning Potentials". In: *The Journal of Chemical Physics* 148.24 (2018), p. 241730.

[368] B. Cheng, R.-R. Griffiths, S. Wengert, C. Kunkel, T. Stenczel, B. Zhu, V. L. Deringer, N. Bernstein, J. T. Margraf, K. Reuter, and G. Csanyi. "Mapping Materials and Molecules". In: *Accounts of Chemical Research* 53.9 (2020), pp. 1981–1991.

[369] S. I. Bityukov, A. V. Maksimushkina, and V. V. Smirnova. "Comparison of Histograms in Physical Research". In: *Nuclear Energy and Technology* 2.2 (2016), pp. 108–113.

[370] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. D. Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, and R. M. Wentzcovitch. "Quantum ESPRESSO: A Modular and Open-Source Software Project for Quantum Simulations of Materials". In: *Journal of Physics: Condensed Matter* 21.39 (2009), p. 395502.

[371] P. Giannozzi, O. Baseggio, P. Bonfà, D. Brunato, R. Car, I. Carnimeo, C. Cavazzoni, S. de Gironcoli, P. Delugas, F. Ferrari Ruffino, A. Ferretti, N. Marzari, I. Timrov, A. Urru, and S. Baroni. "Quantum ESPRESSO toward the Exascale". In: *The Journal of Chemical Physics* 152.15 (2020), p. 154105.

[372] G. Makov and M. C. Payne. "Periodic Boundary Conditions in Ab Initio Calculations". In: *Physical Review B* 51.7 (1995), pp. 4014–4022.

[373] G. J. Martyna and M. E. Tuckerman. "A Reciprocal Space Based Method for Treating Long Range Interactions in Ab Initio and Force-Field-Based Calculations in Clusters". In: *The Journal of Chemical Physics* 110.6 (1999), pp. 2810–2821.

[374] S. Di Pino, E. D. Donkor, V. M. Sánchez, A. Rodriguez, G. Cassone, D. Scherlis, and A. Hassanali. "ZundEig: The Structure of the Proton in Liquid Water from Unsupervised Learning". In: *The Journal of Physical Chemistry B* 127.45 (2023), pp. 9822–9832.

[375] S. Chmiela, H. E. Sauceda, A. Tkatchenko, and K.-R. Müller. "Accurate Molecular Dynamics Enabled by Efficient Physically-Constrained Machine Learning Approaches". In: vol. 968. 2020, pp. 129–154.

[376] C. Shannon. "Communication in the Presence of Noise". In: *Proceedings of the IRE* 37.1 (1949), pp. 10–21.

[377] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, and A. E. Roitberg. "Less Is More: Sampling Chemical Space with Active Learning". In: *The Journal of Chemical Physics* 148.24 (2018), p. 241733.

[378] C. Schran, K. Brezina, and O. Marsalek. "Committee Neural Network Potentials Control Generalization Errors and Enable Active Learning". In: *The Journal of Chemical Physics* 153.10 (2020), p. 104105.

[379] Y. Lysogorskiy, A. Bochkarev, M. Mrovec, and R. Drautz. "Active Learning Strategies for Atomic Cluster Expansion Models". In: *Physical Review Materials* 7.4 (2023), p. 043801.

[380] M. Kulichenko, K. Barros, N. Lubbers, Y. W. Li, R. Messerly, S. Tretiak, J. S. Smith, and B. Nebgen. "Uncertainty-Driven Dynamics for Active Learning of Interatomic Potentials". In: *Nature Computational Science* 3.3 (3 2023), pp. 230–239.

[381] C. van der Oord, M. Sachs, D. Kovacs, C. Ortner, and G. Csanyi. "Hyperactive Learning for Data-Driven Interatomic Potential". In: (2022).

[382] J. Vandermause, S. B. Torrisi, S. Batzner, Y. Xie, L. Sun, A. M. Kolpak, and B. Kozinsky. "On-the-Fly Active Learning of Interpretable Bayesian Force Fields for Atomistic Rare Events". In: *npj Computational Materials* 6.1 (1 2020), pp. 1–11.

[383] F. Ge, R. Wang, C. Qu, P. Zheng, A. Nandi, R. Conte, P. L. Houston, J. M. Bowman, and P. O. Dral. "Tell Machine Learning Potentials What They Are Needed For: Simulation-Oriented Training Exemplified for Glycine". In: *The Journal of Physical Chemistry Letters* 15.16 (2024), pp. 4451–4460.

[384] A. Warshel and M. Levitt. "Theoretical Studies of Enzymic Reactions: Dielectric, Electrostatic and Steric Stabilization of the Carbonium Ion in the Reaction of Lysozyme". In: *Journal of Molecular Biology* 103.2 (1976), pp. 227–249.

[385] H. C. Watanabe, M. Yamada, and Y. Suzuki. "Proton Transfer in Bulk Water Using the Full Adaptive QM/MM Method: Integration of Solute- and Solvent-Adaptive Approaches". In: *Physical Chemistry Chemical Physics* 23.14 (2021), pp. 8344–8360.

[386] S. Pezeshki and H. Lin. "Adaptive-Partitioning QM/MM for Molecular Dynamics Simulations: 4. Proton Hopping in Bulk Water". In: *Journal of Chemical Theory and Computation* 11.6 (2015), pp. 2398–2411.

[387] S. Yan, B. Wang, and H. Lin. "Reshaping the QM Region On-the-Fly: Adaptive-Shape QM/MM Dynamic Simulations of a Hydrated Proton in Bulk Water". In: *Journal of Chemical Theory and Computation* 20.9 (2024), pp. 3462–3472.

[388] R. Guareschi, H. Zulfikri, C. Daday, F. M. Floris, C. Amovilli, B. Mennucci, and C. Filippi. "Introducing QMC/MMpol: Quantum Monte Carlo in Polarizable Force Fields for Excited States". In: *Journal of Chemical Theory and Computation* 12.4 (2016), pp. 1674–1683.

[389] J. Zeng, T. J. Giese, Ş. Ekesan, and D. M. York. "Development of Range-Corrected Deep Learning Potentials for Fast, Accurate Quantum Mechanical/Molecular Mechanical Simulations of Chemical Reactions in Solution". In: *Journal of Chemical Theory and Computation* 17.11 (2021), pp. 6993–7009.

[390] T. J. Giese, J. Zeng, Ş. Ekesan, and D. M. York. "Combined QM/MM, Machine Learning Path Integral Approach to Compute Free Energy Profiles and Kinetic Isotope Effects in RNA Cleavage Reactions". In: *Journal of Chemical Theory and Computation* 18.7 (2022), pp. 4304–4317.

[391] K. Zinovjev. "Electrostatic Embedding of Machine Learning Potentials". In: *Journal of Chemical Theory and Computation* 19.6 (2023), pp. 1888–1897.

[392] A. Hofstetter, L. Böselt, and S. Riniker. "Graph-Convolutional Neural Networks for (QM)ML/MM Molecular Dynamics Simulations". In: *Physical Chemistry Chemical Physics* 24.37 (2022), pp. 22497–22512.

[393] J. A. Semelak, P. Ignacio, K. K. Huddleston, J. Olmos, J. S. Grassano, C. Clemente, S. I. Drusin, M. Marti, M. C. G. Lebrero, A. E. Roitberg, and D. A. Estrin. *ANI Neural Networks Meet Electrostatics: A ML/MM Implementation in Amber*. 2024. URL: `https://chemrxiv.org/engage/chemrxiv/article-details/66fb5c0bcec5d6c142bd7cec`. Pre-published.